Multivariate Analysis Techniques for Optimal Vision System Design

Sharifzadeh, Sara

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Multivariate Analysis Techniques for Optimal Vision System Design

Sara Sharifzadeh

Kongens Lyngby 2015
PhD-2015-371

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, Building 324,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
Fax +45 4588 1399
compute@compute.dtu.dk
www.compute.dtu.dk PhD-2015-371

Summary (English)

The present thesis considers optimization of the spectral vision systems used for quality inspection of food items. The relationship between food quality, vision based techniques and spectral signature are described. The vision instruments for food analysis as well as datasets of the food items used in this thesis are described. The methodological strategies are outlined including sparse regression and pre-processing based on feature selection and extraction methods, supervised versus unsupervised analysis and linear versus non-linear approaches.

One supervised feature selection algorithm based on the existing sparse regression methods (EN and lasso) and one unsupervised feature selection strategy based on the local maxima of the spectral 1D/2D signals of food items are proposed. In addition, two novel feature extraction and selection strategies are introduced; sparse supervised PCA (SSPCA) and DCT based characterization of the spectral diffused reflectance images for wavelength selection and discrimination.

These methods together with some other state-of-the-art statistical and mathematical analysis techniques are applied on datasets of different food items; meat, diaries, fruits and vegetables. These datasets are acquired using three different vision systems; a spectral imaging device called VideometerLab, spectroscopy using spectrometer, and diffused reflectance imaging systems called Static Light Scattering (SLS).

These analyses result in significant reduction in the number of required wavelengths and simplification of the design of practical vision systems.

Summary (Danish)

Nærværende afhandling betragter optimering af spektrale visionsystemer, som anvendes til kvalitetskontrol af fødevarer.

Forholdende mellem fødevarekvalitet, vision baserede teknikker og spektrale signaturer er beskrevet. Visionen instrumenter til fødevareanalyse samt de datasæt af fødevarer, der anvendes i denne afhandling, er beskrevet. De metodologiske strategier beskrives og inkluderer såkaldt sparse regression og præ-processering baseret på feature udvælgelse og feature ekstraktionsmetoder, supervised versus un-supervised analyse, samt lineær versus ikke-lineære metoder.

Baseret på eksisterende sparse regressionsmetoder (Elastic Net og lasso) foreslås en superviseret feature udvælgelses metode. Baseret på lokale maksima af de spektrale 1D/2D signaler af fødevarer foreslås en un-supervised feature udvælgelses algoritme. Desuden introduceres to nye feature udvælgelses- og selektions-strategier; nemlig dels sparse supervised PCA (SSPCA), dels en DCT baseret karakterisering af spektralt diffuse reflektans billeder, som anvendes til bølgelængde udvælgelse samt bølgelængde diskrimination.

Disse metoder bliver sammen med andre state-of-the-art statistiske og matematisk analyse teknikker anvendt på datasæt fra forskellige typer fødevaredata; kød, fødevare-dagbøger, samt frugt og grøntsager. Disse datasæt er fremkommet ved hjælp af tre forskellige visionsystemer; en spektral imaging enhed kaldet VideometerLab, spektroskopi ved hjælp af et spektrometer, og endelig diffust reflektans billeddannende systemer kaldet Static Light Scattering (SLS).

Disse analyser resulterer i en væsentlig reduktion af antallet af nødvendige bølgelængder og leder til en forenkling af udformningen af praktiske visionsystemer.

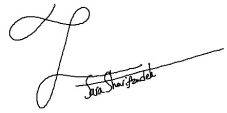
Preface

This thesis was prepared at the Department of Applied Mathematics and Computer Science, Technical University of Denmark in the Data Analysis Section in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering.

The thesis deals with sparse multivariate analysis and dimension reduction techniques for optimal vision systems design. The main focus of this thesis is to reduce the number of wavelengths of the spectral signals and images of food items. Sparse regression methods as well as feature selection and extraction techniques are employed and proposed.

The thesis consists of an introduction to the field of research, the basic materials and methods utilized for the main focus in the thesis and a collection of seven research papers written during the period June 2011- May 2015, and elsewhere published or is under review and one unpublished technical report.

Lyngby, 24-May-2015



Sara Sharifzadeh

Acknowledgements

I would like to thank my supervisors Prof. Line H. Clemmensen and Prof. Bjarne K. Ersbøll for accepting me as a Ph.D. student and guiding me during my study. They supported me through the PhD work and helped me to provide data sets, that was a challenge during my PhD. They patiently answered to all my technical questions and helped me to deepen my knowledge and offered me very useful feedback on my work.

I would like to thank the Danish Strategic Council for funding this project that is part of the projects at the Center for Imaging Food Quality (CIFQ).

I also would like to thank all my friends and colleagues at Data Analysis Section for good atmosphere specially our former and current secretaries Christina H. Nexo and Sladjana E. Pedersen for their help and support. Thanks to all CIFQ industrial and academic parteners, PhD students and colleagues. Thanks to all co-authors of my papers; Particularly, I would like to thank Claus Borggaard at Danish Meat Research Institute (DMRI) for data and technical support and Prof. Hanne Løje at DTU Food for good collaboration, technical and data support. I also thank Jacob L. Skytte for successful collaborations and discussions at CIFQ.

Likewise, I would like to thank Prof. Ali Ghodsi at the Department of Statistics and Actuarial Science , University of Waterloo for hosting me for three months and good collaboration.

I wish to gratefully thank my mother and father and my sisters, Roja and Hilda for their support and encouragement. It was not easy to be so far and be

focused on my research work without them accepting the distance and keeping the connections.

Last but not least, I would like to thank my husband Ehsan and my little son Ryan for their own ways of help, support, kindness and love.

Thank you,
Sara Sharifzadeh

May 2015

Contributions

List of contributions included in this thesis

- **Paper A.** Sara Sharifzadeha, Line Clemmensen, Claus Borggaard, Susanne Støier, Bjarne K. Ersbøll, *Supervised feature selection for linear and non-linear regression of L*a*b* color from multispectral images of meat,* Engineering Application of Artificial Intelligence, vol. 27, p. 211-227, 2014.
- **Paper B.** Sara Sharifzadeh, Jacob Lercke Skytte, Line H. Clemmensen, Bjarne K. Ersbøll, *DCT-Based Characterization of Milk Products Using Diffuse Reflectance Images,* IEEE 18th International Conference on Digital Signal Processing, p. W3C-6, 2013.
- **Paper C.** Sara Sharifzadeh, Ali Ghodsi, Line H. Clemmensen, Bjarne K. Ersbøll, *Sparse supervised principal component analysis (SSPCA) for dimension reduction and variable selection*, Under review.
- **Paper D.** Sara Sharifzadeh, Bjarne K. Ersbøll, Line H. Clemmensen, *An unsupervised strategy for characterization of VIS- NIR spectral signals of food products based on local extrema,* Technical Report.
- **Paper E.** Mabel V Martinez Vega, Sara Sharifzadeh, Dvoralai Wulfsohn, Thomas Skov, Line Harder Clemmensen and Torben B Toldam-Andersen, *A sampling approach for predicting the eating quality of apples using visible-Near infrared spectroscopy*, Journal of Science of Food and Agricalture, vol. 93, p. 3710-3719, 2013.
- **Paper F.** Sara Sharifzadeh, Line Harder Clemmensen, Hanne Løje, Bjarne K. Ersbøll, *Statistical quality assessment of pre-fried carrots using multi-*

spectral imaging, Image Analysis, Springer, Presented at: 18^{th} Scandinavian Conference on Image Analysis, p. 620-629, 2013.

- **Paper G.** Sara Sharifzadeh, Mabel V Martinez Vega, Line H. Clemmensen, Bjarne K. Ersbøll, *Optimal vision system design for characterization of apples using UV/VIS/NIR spectroscopy data*, IEEE 20^{th} International Conference on Systems, Signals and Image Processing, p. 11 - 14, 2013.

- **Paper H.** Sara Sharifzadeha, Hanne Løje, Line Clemmensen, Grethe Hyldig, Bjarne K. Ersbøll, *Sensory quality prediction using multispectral imaging*, Submitted.

List of contributions not included in this thesis

- Sara Sharifzadeh, Jacob L. Skytte, Otto H. A. Nielsen, Bjarne K. Ersbøll, Line K. H. Clemmensen, *Regression and Sparse Regression Methods for Viscosity Estimation of Acid Milk From it's SLS Features*, IEEE 19^{th} International Conference on Systems, Signals and Image Processing, p. 58-61, 2012.

- Philip J. Sassene, Jacob L. Skytte, Sara Sharfizadeh, Line K. H. Clemmensen, Huiling Mu, Thomas Rades, Anette Müllertz, *Evaluation of Off-line Intestinal In Vitro Lipolysis of Lipid-based Drug Delivery Systems (Lb-DDS) by Diffuse Reflectance Imaging (DRI)*, 9^{th} World Meeting on Pharmaceutics, Biopharmaceutics and Pharmaceutical Technology: March, 2014, Accepted.

Contents

Chapter 1

Introduction

The increased expectation for manufacturing high-quality safe food products necessitates accurate, fast, and objective quality determination in the highly competitive food industry. Monitoring the quality of food products using vision based systems has gained a lot of attention in the food industry recently. Reviewing the research literature shows the application of this technology on varieties of food items and the vast research work that has been accomplished in this area. This include different types of meat, diaries, fruits, and vegetables as well as any other types of food items such as chocolate.

The current thesis is part of the projects at the Center for Imaging Food Quality (CIFQ). The main focus of the projects in this center is to employ imaging technologies for quality monitoring of food products. The benefit of these systems is that they are contact-less and non-invasive while giving useful information about the surface and sub-surface of food items with out making any contamination. There are different partner in this project from industry including Videometer A/S, NKT Photonics A/S, Arla Foods, Danish Meat Research Institute and Dupont Nutritional Bio-science.

This thesis involves applying and developing data analysis techniques for optimization of vision based systems used for quality assessment of food items. As a result, the costs of developing such systems will be reduced and they become simpler. In addition, the statistical prediction models built on the related food

data will become more accurate. This thesis is not focused on data acquisition and collection activities and the data sets used in this thesis was provided mainly from either CIFQ partners or other collaborations with DTU Food and University of Copenhagen, Department of Plant and Environmental Sciences and Department of Food Science. However, having insight about the imaging systems used for data preparation and understanding the data characteristics and the desired quality parameters is important for finding the best analytical solutions. Therefore, relevant information about the vision instruments used for data acquisition as well as food data sets specifications are provided and presented in this thesis.

The structure of this chapter is as follows; first the relationship between food quality and vision systems is explained in section 1.1. Then, the concept of spectral signature is described in section 1.2. Section 1.3 is about the main aims of this thesis. Finally there are some reading guidelines for this thesis in section 1.4.

1.1 Food Quality and vision based techniques

Quality is a general term. In the case of food items, quality is usually measured based on visual properties such as color, texture, and chemical composition and physical properties such as fat or protein level, water content, firmness etc. It can be evaluated by a descriptive term or human sensory attribute such as surface browning and water content of biscuits (Dissing, 2011), color and texture of vegetables (Løkke et al., 2013) and tenderness of a piece of meat (Kamruzzaman et al., 2013) but also more quantitative measures of ingredient proportions such as fat size distributions in milk (Cabassi et al., 2013) or sugar content of apples (Sánchez et al., 2003). In this thesis, the terms "quality parameter" and "quality attribute" are used alternatively for any of this kind.

The traditional quality assurance methods used in the food industry involve human visual inspection and are tedious, laborious, time-consuming, and can be inconsistent (ElMasry and Sun, 2010). Vision based techniques have important privileges over the traditional assessment methods. They are fast, non-invasive and contact-less and result in reproductive quality monitoring methods in the food industry. Additionally, they can be used objectively and automatically on-line.

Vision techniques utilize advanced sensing technologies and instrumentation to evaluate quality and quality-related attributes. While ordinary means such as RGB colour cameras are useful for external quality attributes such as size, shape,

colour, and surface texture, it is difficult to detect internal structures such as fat level or map of water content by relatively simple and traditional imaging means (Du and Sun, 2004; ElMasry and Sun, 2010). For this reason, spectroscopy and spectral imaging techniques have been widely used for quality assessment. They are based on optical properties such as reflectance, transmittance, absorbance or scatter of polychromatic or monochromatic radiation over the ultraviolet (UV), visible (VIS), and near-infrared (NIR) regions of the electromagnetic spectrum (Sun, 2009, 2010).

1.2 Spectral signature and vision system design

The quality parameters of food items affects their optical properties such as reflectance and absorbance acquired by the spectral vision systems such as hyper/multi spectral imaging or spectroscopy (Aguilera, 2005).

On the other hand, due to the difference in their chemical compositions and inherent physical structures, all materials reflect, scatter, absorb and/or emit electromagnetic energy in distinctive patterns at specific wavelengths. This characteristic is called spectral signature or fingerprint. Therefore, spectral signature can be used to uniquely characterize any given object using its spectral signal or image over some ranges of wavelengths (ElMasry and Sun, 2010).

The physical and chemical quality information is determined based on the correlation between the spectral response and a specific quality attribute of a product. However, the spectral signal or image is usually acquired in tens or hundreds of wavelengths. Therefore, the dimensionality of the acquired data might be high. However, not all of the wavelengths carry relevant information to the desired quality parameters and many of them may be irrelevant or noisy. On the other hand, the acquisition and analysis time will increase as the number of bands grows. This limits them to be implemented directly in on-line systems for automated quality evaluation purposes. In practice, the use of one or a small number of bands is preferred. Thus, reducing the number of wavelengths is important and makes the design of the acquisition system easy and economic.

Another problem of spectral data is multicollinearity and high correlation. This problem can be alleviated by multivariate analysis techniques and variable selection strategies (Brereton, 2009). This can improve the predictive power of the calibration model and simplify it by avoiding the redundancies and irrelevant variables.

1.3 Main goals of the thesis

The main goal of this thesis is optimization of the spectral vision systems used for food quality assessment. This can be achieved by reducing the number of required wavelengths using multivariate analysis techniques. For this aim, the following issues are considered and included in this thesis:

- Based on the fact that each food item has its own spectral signature that allows ignoring other irrelevant wavelengths, simplifying the vision systems and also improving the quality prediction models is conducted.
- The existing sparse data analysis and feature selection methods are applied on food data sets with different quality challenges.
- New feature selection and extraction methods are developed and applied on varieties of real scenarios to find the desired quality parameters using the spectral data of food items.

1.4 Reading guidelines

In order to make it easier to read this thesis, some guidelines are presented here.

Theoretical and application parts: This thesis consists of two parts: A theoretical part and an application part. In the theoretical part, first the materials including the vision systems and food data sets used in this thesis are described in section 2. Next, in an introductory chapter (chapter 3) the main methodological considerations for this thesis are outlined. The basic methods used in the papers are explained in chapter 4 and finally, for each theoretical paper, a separate chapter is dedicated (chapter 5 - 8). Part II is about the application of some of the methods explained in part I and for each paper that is about the application of the described methods, one chapter is considered (chapter 9 - 12). Finally, we conclude this thesis in chapter 13.

Papers For each paper, an extended abstract is provided as a single chapter in its relevant part. The papers are placed in an appendix. The extended abstracts give an overview of the papers and help the reader to find the details of interest.

Reading Flow This thesis is written so that it can be read from the beginning to the end. As the papers are included in the appendix, the reader does not need to read all of them and can choose upon his interest. In the following, the abbreviations used in this thesis are listed to help the reader.

1.4.1 Abbreviation

ANN Artificial Neural Networks

CCD - Charge Coupled Device

CIFQ - Center for Imaging Food Quality

DCT - Discrete Cosine Transform

DPA - Discrimination Power Analysis

EN - Elastic Net

FL - Fused lasso

HSIC - Hilbert - Schmidt Independence Criterion

LAR - Least angle regression

Lasso - Least Absolute Shrinkage and Selection Operator

LED - Light Emitting Diode

LOOCV - Leave-One-Out Cross Validation

MAP - Maximum a Posteriory

MHT - Multiple Hypothesis Testing

NIR - Near Infrared

PC - Principal Component

PCA - Principal Component Analysis

PLS - Partial Least Square

RBFANN - Radia Basis Function ANN

ROI - Region of Interest

SLS - Static Light Scattering

SPCA - Sparse Principal Component Analysis

SPLS - Sparse Partial Least Square

SSC - Solvable Solid Content

SSPCA - Sparse Supervised Principal Component Analysis

SVD - Singular Value Decomposition

SVM - Support Vector Machine

UV - Ultra Violet

VIS - Visible

Part I

Materials and Methods

Chapter 2

Vision systems and spectral data materials

The data sets used in this thesis were formed using different imaging techniques. In this chapter, first the vision systems are described. Then, the data sets and their related challenges are introduced.

2.1 Vision instruments for food analysis

In the first step of the work, a vision system acquires the digital signal or image of a food item. The signal/image is the reflectance of the light illuminated to a point/surface of the object and sensed by the vision device. The light might be illuminated in several hundreds of wavelengths depending on the spectral range (visible (VIS), near infra red (NIR)) and resolution of the vision instrument. The VIS covers 380-750 nm and the NIR covers 780-2500 nm. Sometimes, the ultra violet (UV) bands which are below 380 nm are also used. The VIS mainly shows the visual and physical characteristics such as color and texture while the NIR regimes signal can be influenced by features that are correlated to the chemical characteristics such as rheology, particle size and etc. Figure 2.1 shows the position of these bands in the frequency spectrum.

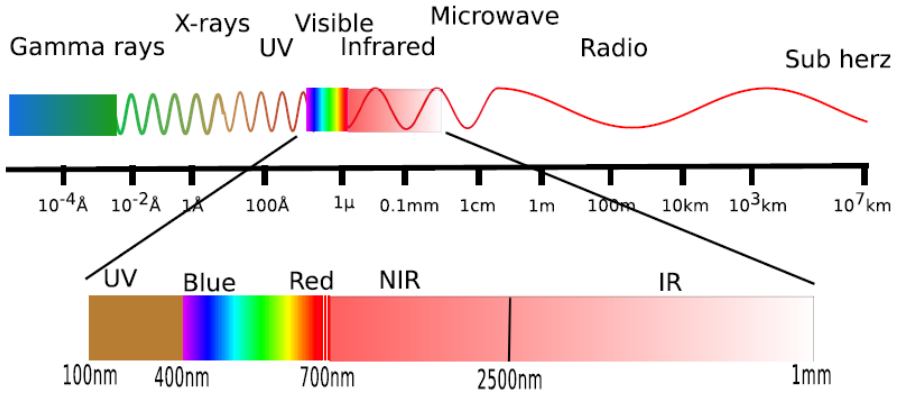


Figure 2.1: Illustration of the position of UV-VIS-NIR regions in the frequency spectrum (Dissing, 2011)

There are different imaging devices and depending on the food items physical and chemical characteristics and desired quality parameters, the most suitable device can be chosen.

2.1.1 Colorimeter

Colorimeters are traditional instruments for measurements of color in different color spaces such as $L^*a^*b^*$ or XYZ in the food industry. They provide a quantitative measurement in a similar way to the human eye (Wu and Sun, 2013; Balaban and Odabasi, 2006). The minolta chromameter and Hunter Lab are examples of colorimeters that are used for food items. However, such traditional instrumental measurements can only measure the surface of a sample that is uniform and rather small (Balaban and Odabasi, 2006). Hence, they cannot completely represent the surface characteristics especially when it is non-uniform and highly textured. Figure 2.2 shows a colorimeter device.

2.1.2 CCD Camera

One of the conventional imaging techniques is based on the analysis of RGB images captured by a Charged Coupled Device (CCD) camera (Pallottino et al., 2010; Larsen et al., 2014). Similar to the human eye, a CCD is capable of absorbing photons in the visible area of the electro magnetic spectrum depicted in

Figure 2.2: A CR-400 chroma meter from Konika Minolta (konicaminolta, 2015)

.

figure 2.1. It has similar filters to the human eye filters or cones; β , γ and ρ. In the standard RGB image, there are three channels; Red (650 nm), Green (510 nm) and blue (475 nm) (see figure 2.3). A CCD converts the electro magnetic spectrum to electrical impulses to be interpreted as numerical values. The common CCD basically integrates all the incoming light in the visible area giving rise to monochromatic images (Dissing, 2011).

The external view and visual characteristics such as color and texture of a food item can be extracted using an RGB image. However, the internal view, that is important for some quality parameters such as chemical components, can not be screened. In addition, it is inefficient in the case of objects of similar colour (ElMasry and Sun, 2010). Figure 2.5(a) shows an RGB image of some wok-fried celeriac.

2.1.3 Spectrophotometer

Another widely used method is spectroscopy. Usually in this method, a spectrophotometer is used to illuminate light into a small surface area in hundreds of bands of electromagnetic spectrum. The interaction of the electromagnetic radiation with atoms and molecules of the object under study that creates optical properties such as reflectance or absorption are analyzed (ElMasry and Sun, 2010). This analysis helps to qualify and quantify chemical and physical information contained within the wavelength spectrum based on the fact that, certain materials have unique fingerprints or spectral signatures in the electromagnetic spectrum. Therefore, it is possible to identify the chemical composition of a food item (that can be related to its quality) based on this. For example, all biological substances contain thousands of C—H (such as organic compounds and petroleum derivatives), O—H (such as moisture, carbohydrate and fat)

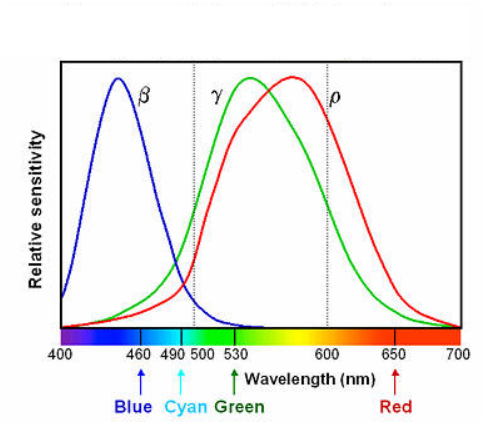


Figure 2.3: Spectral sensitivity of the human eye to color roughly matches to the RGB images

Figure 2.4: View of a spectrophotometer measuring the light diffusely reflected from a fruit (Bernd Herold, 2008)

and N—H (such as proteins and amino acids) molecular bonds. The bonds of organic molecules change their vibration response energy when irradiated by NIR frequencies and exhibit absorption peaks through the spectrum (ElMasry and Sun, 2010). Thus, qualitative and quantitative chemical and physical information is contained within the wavelength spectrum of absorbed energy or reflected energy. Absorbance and reflectance have reverse relation. In figure 2.4 a spectrophotometer that is used for measuring the light diffusion of a fruit is illustrated (Bernd Herold, 2008). Figure 2.5(b) shows the spectroscopy signal of an apple.

However, spectroscopic techniques are point-based so that only the amount of light reflected or transmitted from a specific area of a sample are sensed and do not give information on the spatial distribution of light in the sample. Therefore, they are not suitable for heterogeneous structures.

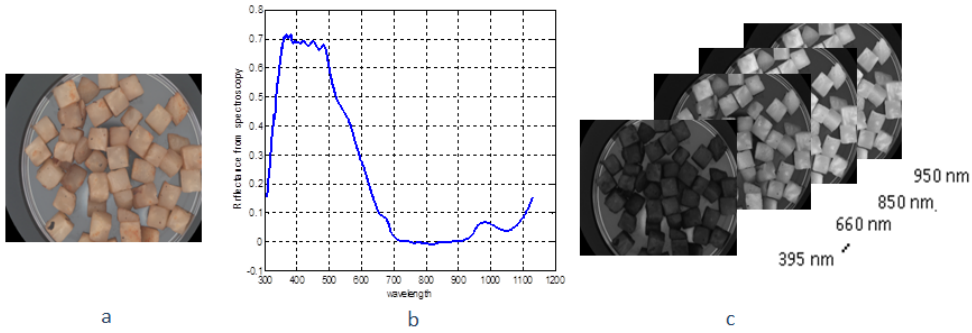


Figure 2.5: (a) an RGB image of some wok-fried celeriac, (b) spectroscopic signal of an apple (c) spectral images of the wok-fried celeriac

2.1.4 Multi/Hyper spectral imaging device

Multi/hyper spectral imaging techniques integrate advantages of conventional image-based techniques and spectroscopy methods. They acquire both spatial and spectral information, forming a series of sub-images, each one representing the intensity distribution at a certain spectral band (see figure 2.5(c)). Therefore, a spectrum is obtained for each pixel in the image of a scene. Depending on the spatial resolution and the structure of the sample under study, the acquired spectra from a Region of Interest (ROI) may show the characteristics of some mixed material instead of a pure spectrum of one singular material (ElMasry and Sun, 2010). Spectral imaging is not suitable for liquids or homogenous samples. A point-wise method such as spectroscopy is sufficient in this case, since the advantage of acquiring 2D images than 1D signals lies in their ability to consider the spatial heterogeneities in samples (ElMasry and Sun, 2010).

There are different methods and devices for capturing spectral images. Most of the spectral images used in this thesis were acquired by VideometerLab. It is a spectral imaging instrument designed for fast and accurate determination of surface color and chemical composition. As shown in figure 2.6(a), it has an integrating sphere or so-called Ulbricht sphere, which has its interior coated with a matt-white coating (Dissing, 2011). The coating, together with the curvature of the sphere provides high diffuse reflectivity for optimal uniform light conditions. There are Light Emitting Diodes (LEDs), positioned side by side in a pattern which distributes the LEDs belonging to each wavelength uniformly around the entire rim. Figure 2.6(b) shows the spectral radiant power distributions of LEDs. As can be seen, the spectral resolution of the LEDs

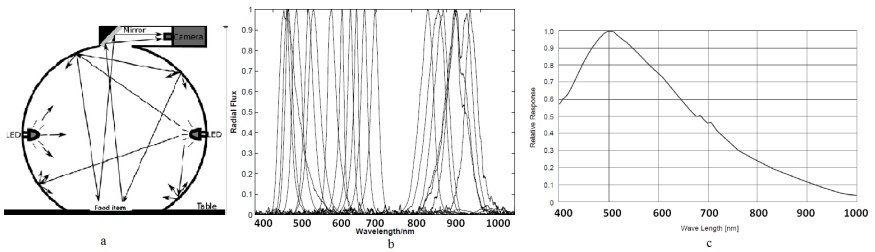


Figure 2.6: (a) Principal set-up of the multispectral imaging system based on integrating (Ulbricht) sphere illumination. The LEDs located in the rim of the sphere ensures narrowband illumination. (b) Normalized spectral power distributions of the LEDs located in the VideometerLab. (c) Spectral sensitivity of the camera mounted in VideometerLab (Dissing, 2011).

are much higher compared to RGB channels shown in figure 2.3. The LEDs strobe successively, each resulting in a monochrome image. These are calibrated radiometrically as well as geometrically to obtain the optimal dynamic range for each LED as well as to minimize distortions in the lens and thereby pixel-correspondence across the spectral bands. In the top of the sphere there is a camera with a sensitivity area corresponding to the desired spectra, as shown in figure 2.6(c). The well defined and diffuse illumination of the optically closed scene aims to avoid shadows and specular reflections. Furthermore, the system has been developed to guarantee the reproducibility of the collected images. This allows for comparative studies of images taken at different times (Gomez et al., 2007). The dense sampling of the electromagnetic spectrum results in acquiring more information about the object and increased ability to distinguish different types of materials and surface chemistry that is useful for food assessment. However, the large amounts of information also creates large amounts of data which needs more complicated processing strategies and more storage capacity (Dissing, 2011).

2.1.5 Static light scattering (SLS)

In this new visioning method, a laser beam is illuminated to the surface of the sample under study at an oblique incident angle of 45° and the resulting spatial distribution of diffuse reflectance is captured using a CCD camera. The laser beam has a high spectral range and resolution ($465-1030nm, \pm 5nm$). Therefore, the system is hyperspectral and the system captures a set of high dynamic range (HDR) images at approximately 2 seconds/wavelength. Figure 2.7 shows the

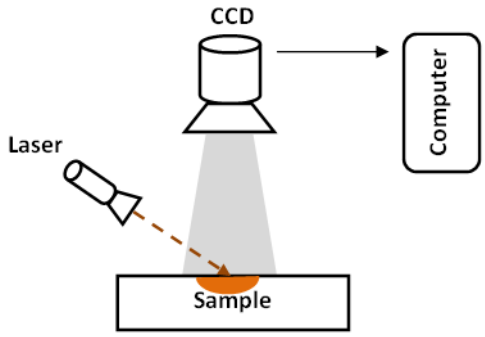


Figure 2.7: General schematic view of the hyperspectral SLS vision system

general schematic of the system. For more details about this system, we refer to (Skytte et al., 2014).

2.2 Food data

In this thesis, different food data sets from varieties of food items have been studied. In this section, they are introduced and the related challenges are described.

2.2.1 Multispectral images of meat

The meat data was provided by the Danish Meat Research Institute. Figure 2.8 shows six different samples of meat from the data set. In this data set, there were multispectral images of different types of turkey, chicken, beef, veal and pork. For each meat sample, multispectral images were acquired at 20 different wavelengths ranging from 430 to 970 nm using a VideometerLab. In addition, the reference values of the $L^*a^*b^*$ color of the samples were available. Two Minolta Chroma Meters CR300 and CR400 were used for that. The aim is to develop a prediction model for $L^*a^*b^*$ values based on the spectral images using a minimum number of bands. Totally, 52 meat samples were digitized. They were divided randomly into training and test sets 25 times. In each data set, the number of training samples were 38. These were used for building the models while the remaining 14 samples were kept as unseen data for the test step.

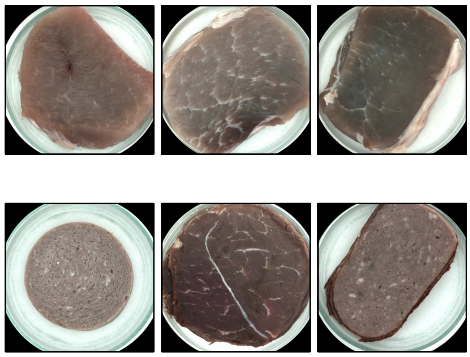


Figure 2.8: Six different meat samples

This data set was used in a paper that is described in chapter 5 and appendix A.

2.2.2 Spectroscopic measurements of apples

Two data sets of apple spectroscopic measurements were analyzed in this thesis. The measurments and data preparation were performed at Copenhagen University.

The first data set consisted of two types of Danish apples, "Aroma" and "Holsteiner Cox". There were 196 middle early season and 219 late season apples. A spectrometer (MOE-1System, Tec5AG, Oberursel,Germany) was used to collect reflectance readings in 1 nm increments within a wavelength range between 400–1130 nm, yielding 731 values per spectrum. In addition the solvable solid content (SSC) and acidity were available for each sample from laboratory measurements. The aim is to develop prediction models for estimation of the SSC and acidity of apples using the reflectance spectra. In addition, different subsampling techniques are used to form training and test sets and their effect on the overall performance of the prediction models is compared. Figure 2.9 shows the raw spectra and its corresponding SSC signal for the apple type.

This data set was used in the paper that is described in chapter 9 and appendix E.

The second data set was from an apple cultivar called “Rajka”. Spectroscopic

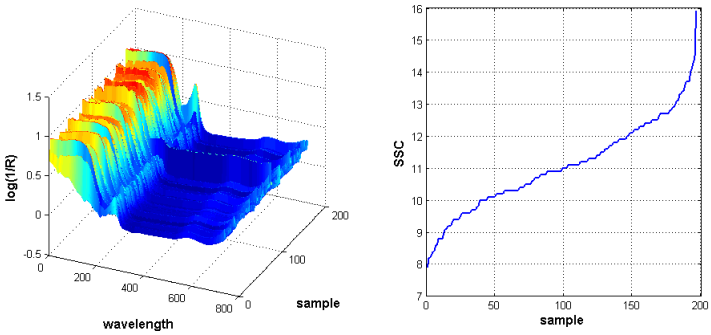


Figure 2.9: Raw spectral patterns recorded in the VIS–NIR region 400–1100nm of apple type 'Aroma'. The absorption is shown which is $\log(\frac{1}{\text{Reflectance}})$ (left) and the corresponding measured SSC (right)

measurements were performed on both sides, exposed and non-exposed to the sun and the average results were considered. The measurements were performed for a total of 825 wavelengths (306-1130 nm) with 1 nm resolution. There were 185 data points (apple samples) in total. In addition, the SSC (%Brix) and the firmness (N) values for each apple were available from laboratory measurements. Figure 2.10 shows the spectroscopic data in UV/VIS and NIR wavelengths as well as the corresponding sorted SSC and firmness signals. The aim of the analysis is to find the proper set of wavelengths carrying relevant information for prediction of SSC and firmness using the spectroscopic measurements. This was done using the sparse regression methods and two model selection strategies. The relation between the choice of statistical methods and the design of the vision setups was also determined.

This data set was used in the papers that are described in chapters 7,8,11and appendixes C, Dand G.

2.2.3 Multispectral images of vegetables

This data set was provided by DTU Food and National Food Institute. The vegetable data set consists of multispectral images of two types of wok-fried vegetables, carrot and celeriac. The vegetables were cut into cubes of size approximately $0.5cm^2$. Two batches of each type were used and there were two replicate samples in each batch. In a pilot plant, the raw products were stir-fried using a special frying machine; "the continuous wok" (Adler-Nissen, 2007).

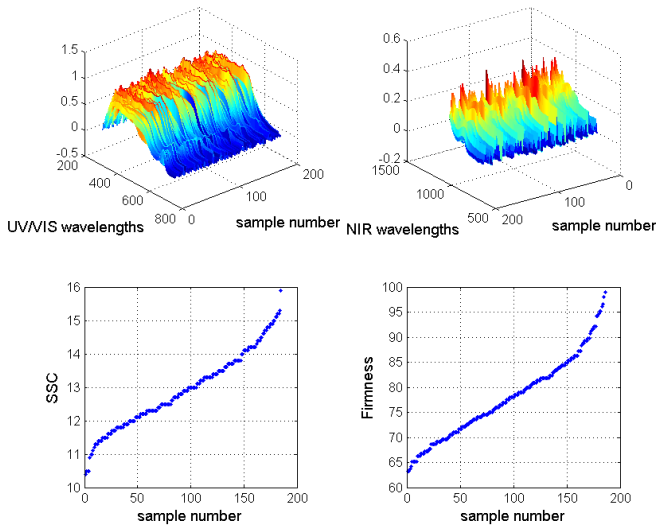


Figure 2.10: The UV/VIS and NIR wavelengths spectra of Rajka apple and the corresponding SSC and firmness signals.

After frying and cooling, the products were packed and frozen to $-30°C$ and after about 60 days of freezing, the bags were removed from the freezer, thawed and kept for up to 14 days at $+5°C$ in a refrigerator. On each day of analysis (days 2, 5, 8, 11 and 14), two polyethylene bags were taken out of the refrigerator and digitized using a VideometerLab. Multispectral images were captured at 19 different wavelengths ranging from 430 to 970 nm. Figure 2.11 shows multispectral images of a carrot sample in a petri dish.

After measurement with the VideometerLab, the samples were reheated and served for a sensory panel of six assessors. The sensory evaluations were performed in a sensory lab under normal daylight and at ambient temperature. At the sensory assessment, appearance, smell, taste and texture were assessed. Each attribute was given a score between zero and two or three demerit points. In addition, an expert assessor score was developed based on the agreement of the 6 assessors on each vegetable sample.

First, the multispectral images of carrot were analyzed to detect any significant changes over the days of storage. The results of this analysis is described in the paper presented in chapter 10 and appendix F.

In the second step, a similar analysis was repeated for celeriac spectral data.

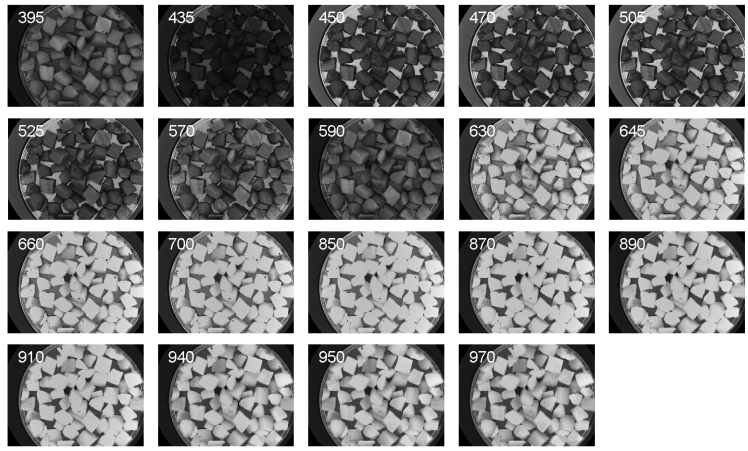


Figure 2.11: Spectral images of carrot at 19 wavelengths. The center wavelength in nanometer is given.

Besides that, the spectral as well as the sensory data of both vegetables were used to develop prediction models for estimation of the sensory attributes using the spectral data. The paper of this work is described in chapter 12 and appendix H.

2.2.4 Diffuse reflectance spectral images of diary products

Two data sets of milk and diary products were provided from CIFQ (Skytte et al., 2014).

One of the analyzed SLS data sets consisted of spectral diffuse reflectance images of eight dairy products including milk and yogurt categories. There were 5 samples available per product (40 samples in total) and the laser was illuminated in 55 wavelengths (460-1000 nm). The products differed from each other in terms of fat and viscosity level. The analysis was performed to characterize and discriminate them using their optical features that represent their chemical, physical and structural differences. Figure 2.12 shows diffuse reflectance images of two different samples. This data set was used in the paper that is described in chapter 6 and appendix B.

The second milk data set consisted of diffuse reflectance images of milk during fermentation process in the controlled condition for fat, temperature and protein

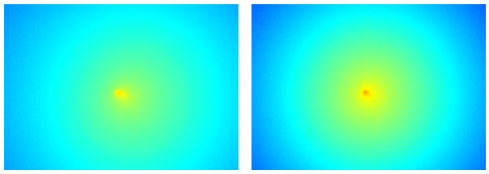


Figure 2.12: Diffuse reflectance images of (left) milk (%1.5) and (right) yogurt (%3.5)

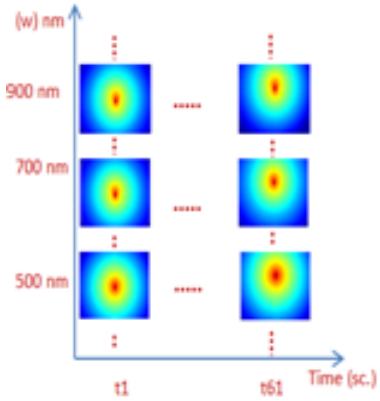


Figure 2.13: The spectro-temporal image set of milk fermentation process (Red corresponds to high pixel intensity and blue corresponds to low pixel intensity)

factors. In fermentation process, every 6 minute the hyperspectral imaging was performed in 57 wavelengths (480-1040 nm). This resulted in a spectro-temporal image set. Figure 2.13 shows the spectro-temporal map of figures obtained during the fermentation process. The process begins with a milk structure at t_1, and ends with a yogurt structure at t_{61}. The experiments were repeated 8 times and in each round, the fat, protein and temperature level was controlled in low or high level, forming a total of 2^3 combinations. In addition, three experiments were conducted so that, all of the factors were in medium level. This data set was used in the technical report that is described in chapter 8 and appendix D. The samples were used for classification into one of the three levels of fat contents using a minimum number of wavelengths and time indexes to simplify the vision set-up and the reduce the complexity of the practical experiments. In this work, the protein and temperature information weren't used. A complete description about this can be found in (Skytte et al., 2014).

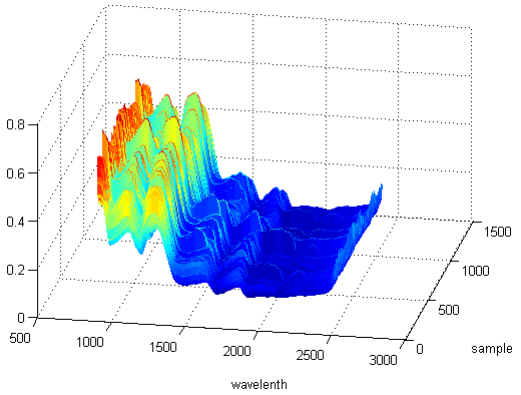


Figure 2.14: The spectral data of 1042 fish pelets in 256 wavelengths

2.2.5 Hyperspectral images of aquaculture feed pellets (NIR)

This data set was provided for a previous PhD work at DTU COMPUTE (Ljungqvist et al., 2012). The data set consists of hyperspectral images of aquaculture feed pellets in the spectral range of 970-2500 nm in a step size of 6.3 nm, resulting in 256 spectral bands in the NIR range captured by a Specim vision system. The fill condition was used where there was white light in the background. The pellets used were coated with five different concentrations of added synthetic astaxanthin (0, 20, 40, 60, 80 ppm). This data set was used in (Ljungqvist et al., 2012). The aim of the study was to investigate the possibility of predicting the concentration level of synthetic astaxanthin coating of feed pellets by NIR hyperspectral image analysis and to distinguish the important spectral features. Figure 2.14 shows a 3D visualization of this spectral data. This data set was used as one of the examined data set in the technical report described in chapter 8 and appendix D.

Chapter 3

Introduction to Methodology

This chapter gives an introduction about the topic of this thesis: multivariate analysis techniques for optimal vision system design. It describes the general strategies and approaches considered in this thesis.

3.1 Methodological strategies

Two different methodological approaches are used in this thesis. One is to employ or develop analytical solutions based on the type of data sets at hand and their related challenges. Such solutions also align with the aims of this thesis. The second approach is to develop solid analytical methods in the context of this thesis goals and test them on different data sets of food items.

Most of the challenges related to the data sets used in this thesis are prediction of a desired quality parameters or characterization and discrimination.

On the other hand, as explained in section 1.2, the quality of each food item is correlated into some of the electromagnetic wavelengths and reducing the number of wavelengths can improve the analysis results.

For the analysis of the data sets mainly two types of analytical methods are employed; sparse regression methods and pre-processing for feature selection or extraction followed by regression or discrimination.

3.1.1 Benefits of a sparse prediction model

In sparse regression methods, a regularization term is added to the loss function of the model that penalizes the complexity of the model. This reduces the variance of the model in price of a small increase in bias. The most correlated variables to the response remain in the model and the other variable's coefficients are shrunken toward zero. This improves the performance of the model. In addition, the resulting model can be interpreted easier.

3.1.2 Benefits of dimension reduction (feature selection and extraction)

For feature selection, an objective function should be maximized or minimized and for feature extraction, the features might be transformed into a new space and some of the features are ignored as they are irrelevant and redundant and do not contain useful information (Clemmensen, 2010). Reduction in the number of variables helps to build a simpler model with less variance. The interpretation also improves and the performance increases.

3.2 Supervised versus unsupervised analysis

In machine learning, a function fitting paradigm that involves learning through a teacher is a supervised learning process. It requires training observations including both input values or predictor variables x_i and outputs or response variables y_i. Using both of these observations, the learning algorithm modifies its output $\hat{f}(x_i) = \hat{y}_i$ based on a loss function $L(y, \hat{y})$ such as the difference $y_i - \hat{f}(x_i)$ between the original and the generated outputs. This process is known as learning by example (Hastie et al., 2009). At the end of learning process, the expectation is that the estimated and real outputs be close enough to each other so that, the algorithm can be used for future inputs likely to be seen in training step.

In contrast, in unsupervised learning or "learning without a teacher", the ob-

servations only consists of the input variables x_i and the outputs y_i are not available. The dimension of X is sometimes much higher than in supervised learning, and the properties of interest are often more complicated (Hastie et al., 2009). These methods look for some patterns in data described by some criterion such as maximum variance, maximum correlation or minimum distance (Clemmensen, 2010). Different statistical tools may be used for analysis such as Gaussian mixtures, clustering, multidimensional scaling, principal component analysis (PCA) and etc. It is difficult to ascertain the validity of inferences drawn from the output of most unsupervised learning algorithms.

Most of the analysis that is performed in this thesis are supervised. However, one unsupervised feature selection method is proposed for the analysis of food items based on local maxima of spectral data. It is explained in chapter 8.

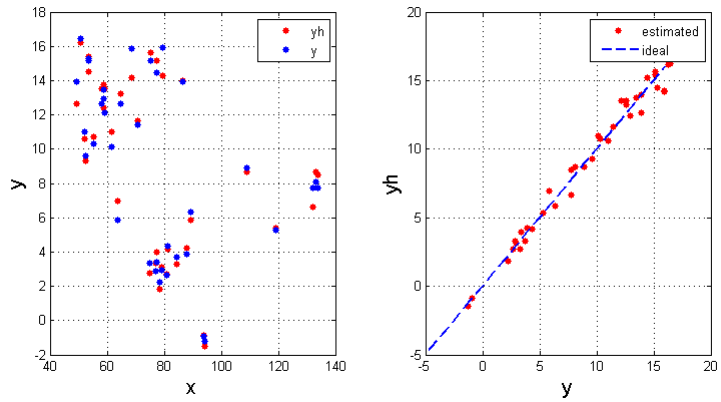
3.3 Linear verses non-linear analysis

In a linear analysis, the relationship between the input or predictor variables x_i and the output or response y_i is modeled as a linear function. For example, a simple linear regression model is as follows:

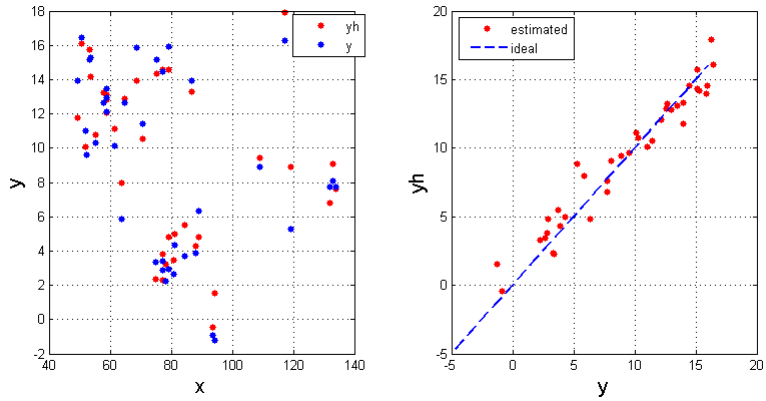
$$\hat{Y} = \beta X + \epsilon = \hat{\beta}_0 + \sum_{j=/}^{P} X_j \hat{\beta}_j + \epsilon \tag{3.1}$$

where $X^T = (X_{1,}, X_2, ..., X_P)$ is the input and the output Y is predicted via this model. In this model, β is an unknown parameter and can be computed by different methods which will be explained in more details in section 4.1.1. $\hat{\beta}_0$ is intercept or bias and ϵ is the residual error. Viewing as a function over the p-dimensional input space, it is linear and in the $1 + P$ dimensional space, $(X, \hat{Y})$ represents a hyperplane.

In contrast, some analysis methods model the relation between input and output as a non-linear function. Examples are the family of artificial neural networks (ANN) or kernel based methods like support vector machine (SVM). In ANN, the main idea is to extract a linear combinations of the inputs as derived features, and then model the target as a non-linear function of these features (Hastie et al., 2009). Depending on a classification/regression task, SVM produces non-linear boundaries/estimation by constructing a linear boundary/estimation in a large, transformed version of the feature space. In this thesis both ANN and SVM were used for regression. They will be described in more detail in section 4.1.2.



(a)



(b)

Figure 3.1: (a-left) The scatter plot of the original as well as estimated a^* color component of meat (using the linear OLS method) versus the input signal in one NIR band (in each input image, the average ROI was considered). (a-right) the estimated a^* by OLS versus the original color component. (b-left and right), showing the same as (a) using a non-linear RBFANN.

The choice of a linear or non-linear method depends on the behavior of the data. It is useful to compare the performance of the two strategies for making an appropriate decision. As an example, figure 3.1 illustrates the regression results for prediction of the a^* color component of meat data described in section 2.2.1. In this figure, y is the a^* color component that is estimated using the multispectral images of meat. In figure 3.1a in left side, both the original as well as the estimated a^* color component by the OLS method are plotted versus the input from one NIR wavelength. A linear trend between the input and output can be observed. The estimated samples are close to the original ones in most cases. In the right side, the estimated a^* color component verses the original one is shown. As can be seen, they closely follow the ideal line. Figure 3.1b shows the same for the non-linear RBFANN method. This regression method is explained more in section 4.1.2.1. Compared to the OLS, the results obtained by RBFANN is less accurate. That shows the effect of the type of relation between input and output and data behavior, on the selection of the prediction strategy.

Chapter 4

Basic Methods

In this chapter, the basic methods that the included papers use them or are based on are described. Some of these methods are also used in the papers included in the application part.

The methods are categorized so that, in the first section just regression methods are described. The second section is about the pre-processing methods. Finally, the model selection strategies used in this thesis are explained.

4.1 Regression methods

In this section, the linear as well as non-linear regression methods used in this thesis are described.

4.1.1 Linear regression methods

4.1.1.1 Ordinary least square (OLS)

One of the popular linear regression methods is OLS. Assuming a P dimensional input vector $X^T = (X_1, X_2, ..., X_p)$, and a real-valued output or response vector Y , a linear regression model as shown in section 3.3 has the form:

$$\hat{Y} = \beta X + \epsilon = \hat{\beta}_0 + \sum_{j=1}^{P} X_j \hat{\beta}_j + \epsilon \tag{4.1}$$

where ϵ is the model error or residual and is assumed to be independent and normally distributed. The most popular and simplest method for computation of the unknown parameter β is the minimization of the residual sum of squares. β_{OLS} includes the intercept parameter β_0 as a 1 is included into x_i.

$$RSS(\beta_{OLS}) = \sum_{i=1}^{N} (y_i - x_i^T \beta_{OLS})^2 \tag{4.2}$$

$RSS(\beta)$ is a quadratic function of the parameters, and hence its minimum always exists. Changing the notation into matrix and taking the derivative with respect to β, a unique solution is obtained. $X^T X$ should be non singular and X is a matrix of N rows showing the observations and $1 + P$ columns of P variables. The first column is of 1s for including the intercept.

$$RSS(\beta_{OLS}) = (Y - X\beta_{OLS})^T (Y - X\beta_{OLS}), \tag{4.3}$$

$$\frac{\partial RSS}{\partial \beta} = X^T (Y - X\beta_{OLS}) = 0 \tag{4.4}$$

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y \tag{4.5}$$

Figure 4.1 shows a geometrical representation of the least squares estimate in a R^3. As can be seen the residual $y - \hat{y}$ is orthogonal to the hyperplane. The minimization of RSS_β results in the β_{OLS} so that the residual vector is orthogonal to this subspace. This orthogonality is expressed in equation 4.4.

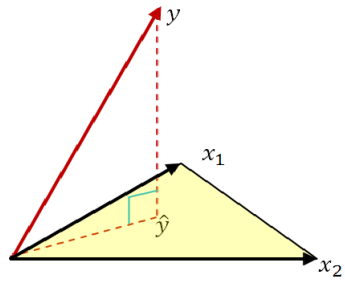


Figure 4.1: Illustration of the N-dimensional geometry of least squares regression with two predictors. $\hat{y}$ is the predicted response and y is the real response. The predicted output is the orthogonal projection of y onto the hyperplane spanned by the input vectors x_1 and x_2.

In many data sets, the columns of X might not be totally independent (which is the case for spectral signals and images) or the number of observations N is smaller than the number of variables P (which happens for many real data). This causes the X not to bet of full-rank. One way to solve this problem is a kind of pre-processing to filter some of the covariates or applying a regularization term. Some related methods in this case are explained in the following sections. For more information we refer to (Hastie et al., 2009).

4.1.1.2 Ridge regression

As stated above, when there are many correlated variables in a data set, the prediction of the variable's regression coefficients is very difficult due to the rank deficiencies. The poorly determined coefficients have high variance. Ridge regression (Hoerl and Kennard, 1970) alleviates this problem so that, the coefficients sizes are penalized by adding a norm two (L_2) regularization constraint to the least square problem.

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{P} \beta_j^2 \right\} \tag{4.6}$$

In this way, the coefficients are shrunken toward zero and $\lambda \geq 0$ controls the amount of shrinkage. It is necessary to normalize the inputs before solving the ridge as it is not equivalent under scaling of the inputs. Therefore, the input matrix X has P (rather than $P + 1$) columns. The matrix form notation is

considered to find β:

$$RSS(\lambda) = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta, \tag{4.7}$$

$$\hat{\beta}_{ridge} = (X^TX + \lambda I)^{-1}X^TY \tag{4.8}$$

where I is the $P \times P$ identity matrix. The ridge regression solution β_{ridge} is again a linear function of y. Always a positive constant is added to the diagonal of X^TX before inversion to avoid singularity.

On the other hand, the singular value decomposition (SVD) of the centered input matrix X helps to find additional insight into the nature of ridge regression (Hastie et al., 2009).

$$X_{N\times P} = UDV^T \tag{4.9}$$

where $U_{N\times P}$ and $V_{P\times P}$ are orthogonal matrices spanning the column and row space of X respectively. D is a $P \times P$ diagonal matrix of singular values and $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$. Applying this for X in ridge regression, we have:

$$X\hat{\beta}_{ridge} = X(X^TX + \lambda I)^{-1}X^Ty = UD(D^2 + \lambda I)^{-1}DU^Ty = \sum_{j=1}^{P} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y, \tag{4.10}$$

where the u_j are the columns of U. Since $\lambda \geq 0$, we have $\frac{d_j^2}{d_j^2+\lambda} \leq 1$. Thus, ridge regression computes the coordinates of y with respect to the orthonormal basis U. For the principal components $z_j = Xv_j = u_jd_j$, the variance $\frac{d_j^2}{N}$ decreases as j increases. From equation 4.10 it can be observed that ridge shrinks the basis vectors or normalized principal components u_j by the factors $\frac{d_j^2}{d_j^2+\lambda}$. Therefore, the coordinates of the basis vector with smaller variance d_j^2 are shrunken more. Figure 4.2 shows a two dimensional data and its corresponding principal components direction. As can be seen, one of the components is larger than the other and therefore, its corresponding direction maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects y onto these components, and then shrinks the coefficients of the low-variance components more than the high-variance

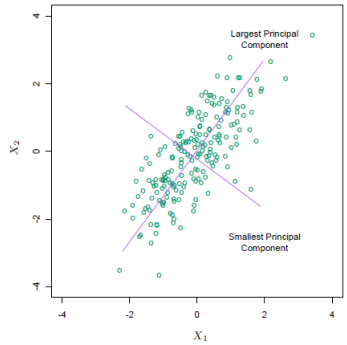


Figure 4.2: Illustration of a two dimensional data and its corresponding principal components. (Hastie et al., 2009)

components. Since in most cases the predictors vary with the response variable, the ridge shrinkage strategy is reasonable. However, it cannot be considered as a general case for all data sets.

4.1.1.3 Lasso

The lasso is a shrinkage method like ridge, but it also has the sparsity nature, which means that some of the coefficients may become zero in shrinkage (Tibshirani, 1994). Instead of a norm two constraint, a norm one constraint (L_1) is applied:

$$\beta_{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{P} |\beta_j| \right\} \tag{4.11}$$

The computation procedure for β_{lasso} is different from ridge. Substitution of (L_2) by an (L_1) constraint makes the solutions non-linear in the y_i, and there is no closed form expression as in ridge regression. A higher λ value means more shrinkage as was also in ridge. However, ridge regression does a proportional shrinkage ($\frac{\hat{\beta}_j}{1+\lambda}$), while Lasso translates each coefficient by a constant factor λ which is called "soft thresholding , $\text{sign}(\hat{\beta}_j)(\left|\hat{\beta}_j\right| - \lambda)_+$. One well known illustrative comparison of ridge and lasso for a two variable case is depicted in figure 4.3. As can be seen, the constraint region for ridge regression is the disk $\beta_1^2 + \beta_2^2 \leq t^2$, while for lasso it is the diamond $|\beta_1| + |\beta_2| \leq t$. The solution is

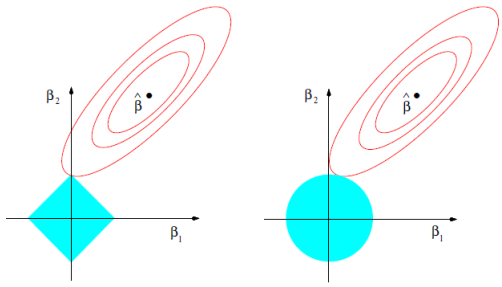


Figure 4.3: Comparison of the error and constraint functions for the lasso (left) and ridge regression (right) (Hastie et al., 2009). The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.

where the elliptical contours of least square hit the constraint region. Since the diamond has corners, if the solution occurs at a corner, then its corresponding parameter become zero. This might also happen when $P > 2$ and explain the sparsity of lasso solution.

Lasso can be calculated using the Least angle regression (LAR) method just by a simple modification. LAR builds a model sequentially, adding one variable at a time. At each step, it identifies the best variable to include in an active set, and then updates the least squares fit. Not all the variables should necessarily be added to the model. At the first step, the coefficient of the variable that is most correlated with the response is moved continuously toward its least squares value (causing its correlation with the evolving residual to decrease in absolute value). As soon as another variable "catches up" in terms of correlation with the residual, the process is paused. The second variable then joins the active set, and their coefficients are moved together in a way that keeps their correlations tied and decreasing. This process is continued until all the variables are in the model, and ends at the full least-squares fit. However, the minimum error for test data may be found in a middle step before all the β coefficients be calculated. In (Hastie et al., 2009), more explanation in this case could be found. In contrast to LAR, if in a lasso path a non-zero coefficient hit zero, its variable should be dropped out from the active set of variables and be treated like other zero coefficients.

4.1.1.4 Elastic net (EN)

Elastic-net is in fact a compromise between lasso and ridge (Zou and Hastie, 2005; Hastie et al., 2009). Looking to the formulation of elastic net in equation 4.12 shows that how each coefficient is calculated as a weighted combination of ridge and lasso. The elastic-net selects variables like the lasso and shrinks together the coefficients of correlated predictors like ridge. This will reduce the variance but at the same time the bias is not as much as it is in lasso.

$$\begin{aligned}\beta_{EN} &= \arg\min_{\beta}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{P}x_{ij}\beta_j)^2+\lambda\sum_{j=1}^{P}((1-\alpha)\left|\beta_j\right|+\alpha\beta_j^2)\right\}\\ &= \arg\min_{\beta}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{P}x_{ij}\beta_j)^2+\lambda_1\sum_{j=1}^{P}(\left|\beta_j\right|)+\lambda_2\sum_{j=1}^{p}(\beta_j^2)\right\}\end{aligned} \tag{4.12}$$

4.1.1.5 Fused lasso (FL)

Fused lasso is a generalized version of lasso that encourages sparsity by means of the (L_1) norm penalty on both regression coefficients and their successive differences (Tibshirani et al., 2005):

$$\beta_{FL} = \arg\min_{\beta}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{P}x_{ij}\beta_j)^2+\lambda_1\sum_{j=1}^{P}\left|\beta_j\right|+\lambda_2\sum_{j=2}^{P}\left|\beta_j-\beta_{j-1}\right|\right\} \tag{4.13}$$

The fused lasso is especially useful for the $N \ll P$ cases, since it sets many coefficients to zero and finds groups of close features. The first penalty term encourages sparsity in the coefficients and the second one encourages sparsity in their differences. Therefore, with this solution, groups of adjacent wavelengths are found.

4.1.1.6 Partial least square (PLS)

PLS is one of the widely used linear regression methods. It uses both Y, and X, for construction of a set of linear combinations of the inputs for regression. Therefore, its solution path is a non-linear function of Y. PLS seeks directions that have high variance and have high correlation with the response. For computation of the regression coefficients β_{PLS}, first successive optimization is performed to calculate $W = (w_1, w_2, ..., w_K)$, so that:

$$w_k = \arg\max_{w} \ (\text{cor}^2(Y, Xw)\text{var}(Xw)) \quad \textbf{s.t.} \ w^T w = 1, \quad w^T \Sigma_{XX} w_j = 0 \tag{4.14}$$

for $j = 1, ..., k-1$, where Σ_{XX} is covariance of X and K is the number of latent components. Then, the latent component matrix $T_{N\times K} = XW$ is computed thereby the response matrix $Y_{N\times q}$ and the predictor matrix $X_{N\times P}$ are decomposed into latent vectors; $Y = TQ^T + F$ and $X = TP^T + E$. $T_{N\times K}$ is a matrix of K linear combinations (scores), $P_{p\times k}$ and $Q_{q\times k}$ are matrices of coefficients (loadings) and $E_{n\times p}$ and $F_{n\times q}$ are matrices of random errors.

One way to solve the PLS problem is using the statistically inspired modification of PLS (SIMPLS) (de Jong, 1993) in which, the k^{th} estimated direction vector $\hat{w}_k$ is found by solving the following optimization problem:

$$\hat{w}_k = \arg\max_{w} \ w^T \sigma_{XY} \sigma_{XY} w \quad \textbf{s.t.} \quad w^T w = 1, \quad w^T \Sigma_{XX} w_j = 0 \ , \tag{4.15}$$

σ_{XY} and Σ_{XX} are the populations covariances of X and Y that can be replaced by the samples covariances (S_{XX}, S_{XY}):

$$w_k = \arg\max_{w} \ w^T X^T Y Y^T X w \quad \textbf{s.t.} \ w^T w = 1, \quad w^T S_{XX} w_j = 0 \tag{4.16}$$

For the details of solution we refer to (de Jong, 1993). Using W, the latent components T and loadings Q are computed. Finally, $\hat{\beta}_{PLS}$ is obtained by $\hat{\beta}_{PLS} = \hat{W}\hat{Q}^T$.

An older solution for PLS is an iterative algorithm as explained in (Hastie et al., 2009).

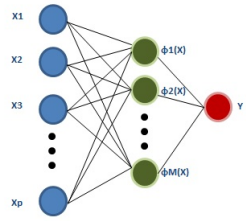


Figure 4.4: The ANN diagram for regression with one hidden layer.

4.1.2 Non-linear regression methods

There are different non-linear methods for regression. In this thesis two of them are used; Artificial Neural Net works (ANN) and Support Vector Machine (SVM) as a kernel-based method.

4.1.2.1 Artificial neural networks (ANN)

The general architecture of a simple ANN for regression with one hidden layer is shown in figure 4.4. First, M linear combinations of the input variables are built and then each combination is transformed using an activation function $h(.)$:

$$\phi_j(X) = h(\Sigma_{i=1}^{i=P} \alpha_{ij} x_k + \alpha_{0j}), j = 1, .., M \tag{4.17}$$

where α_{ij} is the weight parameter and α_{0j} is the bias. Then, the output $\hat{Y}$ is constructed as a linearly weighted combination of the non-linear basis functions $\phi_j(X)$:

$$\hat{Y}(X;\beta) = f\left(\sum_{j=1}^{M} \beta_j \phi_j(X) + \beta_0\right) \tag{4.18}$$

β_j and β_0 are the weight and bias parameters respectively, and $f(.)$ is an activation function which is usually, the identity function in the case of regression (Bishop, 2006).

ANN models are complex and difficult to interpret. Depending on the nature of data, they might result in better performance than linear methods. The number of hidden layers and neurons influence the architecture of an ANN. In addition, the choice of basis function can determine the type of ANN. In the following the three important types that is used in this thesis are described:

feed-forward ANN One widely used ANN is the single hidden layer feed-forward ANN which uses a sigmoid basis function:

$$\phi_j(X) = \sigma_j(X) = \frac{1}{1 + exp(-S_j X)} \tag{4.19}$$

where, S_j is the scale parameter which controls the activation rate. A large scale may cause hard activation around 0.

Radial Basis Function ANN (RBFNN) RBFNN uses a non-linear RBF based on Euclidean distance or Mahalanobis distance (like a Gaussian kernel function):

$$\phi_j(X) = \rho_j \left(\|X - \mu_j\| \right) \tag{4.20}$$

Where μ_j is the center vector of the j^{th} hidden node and ρ is the distance function. The RBFNN also has one hidden layer.

Parameter estimation The parameters of the ANN models are commonly estimated by minimization of the sum of square function as shown in equation 4.21. The Back Propagation (BP) procedure is used (Hastie et al., 2009) to solve this which is a gradient descent process.

$$E(\beta) = \min \sum_{n=1}^{N} \left\| \hat{Y}(X_n; \beta) - Y \right\|^2 \tag{4.21}$$

BPANN is a well known and widely used network. Although it is a powerful algorithm, it has some drawbacks. One important problem with the error function minimization for complex and flexible models is the over-fitting on training

data and poor generalization. Because a complex model is more flexible in capturing the training data behavior. Other problems are slow convergence and the possibility that the network converges to a local minima.

Due to these problems, different strategies are employed as can be found in the literature (Bishop, 2006, 2003; Hagan et al., 1996). Examples are ANN with Adaptive learning rate and momentum term and different regularization approaches to constrain the parameters.

ANN with Adaptive Learning Rate and Momentum Term Considering the error minimization in Equation 4.21, the gradient $\nabla E(\beta)$ can be obtained by means of back-propagation of errors through the layers. This gradient is used in the family of gradient training algorithms which iteratively form:

$$\beta_{k+1} = \beta_k - \eta_k \nabla E(\beta^k), k = 0, 1, 2, ... \tag{4.22}$$

where β_k is the current weight, $-\eta_k$ is the learning rate and k is the step number and $-\eta_k \nabla E(\beta^k)$ shows the search direction. The BP gradient-based training algorithms minimize the error function using the above gradient decent or steepest descent method with constant, heuristically chosen, learning rate. The learning rate determines how fast a network will learn the relationships between input and output patterns. A smaller value of the learning rate means a slower learning process. In fact, the optimal learning rate changes during the training process, as the algorithm moves across the performance surface. Therefore, the performance of the steepest descent algorithm would improve if the learning rate change during the training process. An adaptive learning rate attempts to keep the learning step size as large as possible while keeping learning stable (Hagan et al., 1996).

The idea about using a momentum BP is to stabilize the weight change and smooth the oscillation in the trajectory. Therefore, a fraction of the previous weight change $\Delta\beta^k$ is considered in updating of the current weights β^{k+1}. Acting like a low-pass filter, momentum allows the network to ignore small local minima in the error surface and slide through them. It also speeds the convergence because, when all weight changes are in the same direction, the momentum amplifies the learning rate.

$$\Delta\beta^{k+1} = \gamma \Delta\beta^k - (1 - \gamma)\eta_k \nabla E(\beta^k), k = 0, 1, 2, ... \tag{4.23}$$

where γ is the momentum coefficient and should be between 0 and 1. This

gives the system a certain amount of inertia since the weight vector will tend to continue moving in the same direction unless opposed by the gradient term.

Regularization of ANN

Feed-Forward ANN Regularization The simplest regularizer is the quadratic in which, a penalty term is added to the error function and penalizes the sum of weights toward zero similar to the regularization of the linear methods. This is called weight decay and is shown in equation 4.24. λ is the regularization ratio which controls the trade-off between fitting the data and generalization of the model.

$$\min \left(\sum_{n=1}^{N} \|\hat{y}(x_n;\beta) - y_n\|^2 + \lambda \sum_{i=1}^{N} \beta_i^2 \right) \tag{4.24}$$

One strong regularization method is the Bayesian regularization that estimates the ANN parameters by a probabilistic approach (Bishop, 2006). Both the model output targets Y and parameters β are characterized as random variables with normal distributions. Then, the Bayesian rule is applied, to calculate their prior and posterior probabilities. Consequently, the predictive distribution of the output is obtained, using the sum and product rules for probabilities as shown in equation 4.25. For more details we refer to (Bishop, 2006, 2003).

$$P(\hat{Y} \mid X, Y_{tr}) = \int P(\hat{Y} \mid X, \beta).P(\beta \mid Y_{tr}) d\beta \tag{4.25}$$

where, Y_{tr} denotes the data used for training the model. The averaging nature of the Bayesian method over many different possible solutions solves the over-fitting problem.

BPANN are sensitive to the number of neurons in their hidden layers. Too few neurons can lead to under fitting and too many neurons can cause over fitting. Therefore, for training of the ANN algorithms it is necessary to loop over the number of hidden nodes for an appropriate choice. In addition, it is useful to restart the network and train from different initial points to avoid falling in a local minima.

RBFNN Regularization For generalization of the RBFNN, the GRNN is used (Specht, 1993). In GRNN, the best prediction with minimum variance is obtained as the conditional mean value of Y_{tr} given X.

$$\hat{Y}(X) = E\langle Y_{tr} \mid X\rangle = \int_{-\infty}^{+\infty} Y_{tr} P(Y_{tr} \mid X) dY_{tr} \tag{4.26}$$

This could be calculated using the joint probability. GRNN uses a non-parametric approach to calculate the joint probability $P(X, Y_{tr})$ by a Gaussian isotropic kernel (Parzen window). The resulting probabilistic output is shown in equation 4.28. The numerator is the sum of the weighted training targets which contribute according to their joint probabilities with the input test sample, to form the output target. The denominator normalizes the solution.

$$\hat{Y}(X) = \frac{\int_{-\infty}^{+\infty} Y_{tr} P(X, Y_{tr}) dY_{tr}}{\int_{-\infty}^{+\infty} P(X, Y_{tr}) dY_{tr}} \tag{4.27}$$

$$\hat{Y}(X) = \frac{\sum_{i=1}^{N} Y_{tr}^{i} exp(-\frac{D_i^2}{2\sigma^2})}{\sum_{i=1}^{N} exp(-\frac{D_i^2}{2\sigma^2})} \tag{4.28}$$

where $D_i = (X - X_{tr}^i)^T (X - X_{tr}^i)$ and Y_{tr}^i, X_{tr}^i are the i^{th}training sample values. σ is the standard deviation of the Gaussian kernel and is called the smoothing parameter. As can be realized from this equation, the contribution weights are in fact the Mahalanobis distance of the test input from the training samples. This means that the closer training samples will contribute more in the prediction of the output target. The smoothing parameter has great effect on the output prediction. With larger σ, more training data will contribute in the target output than with a small σ.

4.1.2.2 Support vector machine (SVM)

SVM can be used for both classification as well as regression problems. There is no assumption about the distribution of the population in this method. In this thesis, it have been used for both purposes. Therefore, both methods will be explained briefly. For more details we refer to (Hastie et al., 2009). In this thesis, the LIBSVM toolbox was used for solving the SVM problem (Chang and Lin, 2011).

SVM for classification Considering a set of N pairs of training samples $(x_1, y_1), (x_2, y_2), ...(x_N, y_N)$, with $x_i \in R^P$ and $y_i \in \{-1, 1\}$ so that, the two classes have overlap in feature space, the decision boundary is defined as a hyperplane (in this case a line, $f(x) = x^T\beta + \beta_0$) that creates the biggest margin $M = \frac{1}{\|\beta\|}$ between the training points of the two classes. SVM defines slack variables for such samples $\xi = \{\xi_1, \xi_2, ..., \xi_N\}$ and for those in the right side $\xi_i = 0$ as shown in figure 4.5. A convex quadratic optimization problem should be solved to compute the parameters of the boundary function:

$$\begin{array}{c} \min_{\beta\beta_0} \frac{1}{2} \|\beta\|^2 + C\Sigma_{i=1}^{N}\xi_i \\ \text{subject to } \xi_i \geq 0, y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \forall i \end{array} \tag{4.29}$$

ξ_i is the proportional amount by which the prediction $f(x_i) = x_i^T\beta + \beta_0$ is on the wrong side of its margin. Hence by bounding the sum $\Sigma\xi_i$, the total proportional amount by which predictions fall on the wrong side of their margin are bounded.

Here we do not explain the details about the solution of this optimization problem which results in finding β and β_0. Using these parameters, the decision function can be written as $\hat{G}(x) = \text{sign}\left[\hat{f}(x)\right] = sign\left[x^T\beta + \beta_0\right]$. More details can be found in (Hastie et al., 2009).

SVM can also be defined based on kernels by enlarging the feature space using basis expansions such as polynomials or splines. That is, class separation with linear boundaries improves in higher dimensional spaces and they become nonlinear boundaries when transferred into the original space. For this aim, a nonlinear basis function $h_m(x), m = 1, 2, ..., M$ is considered. In the next step, the SVM classifier is defined using input features $h(x_i) = (h_1(x_i), h_2(x_i), ..., h_M(x_i))$, $i = 1, ..., N$, thereby a nonlinear function $\hat{f}(x) = h(x)^T\beta + \beta_0$ is created. In order to kernelize this function, the optimization problem and its solution are re-written based on the inner product of the input features. This results in replacement of a kernel instead of the inner product of the transformed feature vectors $h(x_i)$:

$$K(x, x') = \left\langle h(x), h(x') \right\rangle \tag{4.30}$$

K should be a symmetric positive (semi-) definite function such as a polynomial, radial basis or hyperbolic tangent (sigmoid). The new kernelized boundary function is $f(x) = \sum_{i=1}^{N} \hat{\alpha}_i y_i K(x, x_i) + \hat{\beta}_0$ which is used by the decision function $\hat{G}(x) = \text{sign}\left[\hat{f}(x)\right]$.

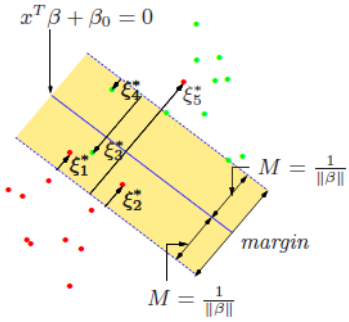


Figure 4.5: Support vector classifiers for a two class problem. The decision boundary is shown by a solid line and the shaded maximal margin is bounded by the broken lines with the width $M = \frac{1}{\|\beta\|}$ on each side. The support vectors are the training samples located inside the margin area and labeled as ξ_j^* (Hastie et al., 2009).

SVM for regression SVM can also be used for regression. Similar to classification, for regression purpose it is also characterized based on a maximum margin algorithm. Given the set of training data $\{(x_1, y_1), ..., (x_N, y_N)\}$, SVM finds a $f(x)$ function that has at most ε deviation from the actual target y. For a linear regression the input feature space is used and for a nonlinear generalization (that we explain here), the features are mapped to an M-dimensional feature space using non-linear basis functions $h(x)$. This is similar to classification. Then, a linear model is constructed in this feature space:

$$f(x, \beta) = \sum_{m=1}^{M} \beta_m h_m(x) + \beta_0 \tag{4.31}$$

To estimate β_m and β_0, the following objective should be minimized:

$$\min_{\beta,\beta_0} H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2} \|\beta_m\|^2 \tag{4.32}$$

$V(.)$ is a loss function called $\varepsilon - sensitive$ defined based on the residual $r = y_i - f(x_i)$:

$$V_\varepsilon(r) = \begin{cases} 0 & if\ |r| < \varepsilon \\ |r| - \varepsilon & \text{otherwise} \end{cases} \tag{4.33}$$

The function ignores errors of size less than ε . There is a similarity between

this function and the idea used in support vector classification, where points on the correct side of the decision boundary and far away from it, are ignored in the optimization. In regression, these low error points are the ones with small residuals (Hastie et al., 2009).

The second term in equation 4.32 controls the complexity level of the model. This optimization leads to a kernel based solution:

$$\hat{f}(x) = h(x)^T \hat{\beta} = \sum_{i=1}^{N} \alpha_i K(x, x_i), \hat{\alpha} = (HH^T + \lambda I)^{-1} Y \tag{4.34}$$

where $K(x, x_i) = \sum_{m=1}^{M} h_m(x) h_m(x_i)$. Then similar to the support vector machine, there is no need to specify or evaluate the large set of functions $h_1(x), h_2(x), ..., h_M(x)$. Only the inner product kernel need be evaluated. For more information, we refer to (Hastie et al., 2009).

4.2 Pre-processing methods

In data analysis, pre-processing is used to prepare data before the analysis for example to remove the noisy, redundant and irrelevant information from data. This leads to a reduction in the dimensionality of data that improves the generalization and training time. There are different pre-processing approaches. In this thesis some feature extraction and feature selection methods are used.

4.2.1 Feature extraction

Feature extraction is a dimension reduction strategy. When the number of spectral variables are high, it is more likely that some of the variables be redundant or not representative enough to be used directly. In feature extraction approach, features are projected into a new space with lower dimensionality (Alelyani et al., 2013). The extracted features are expected to contain relevant information from the input data, so that they can result in better performance than the initial data. Feature extraction can be performed as an unsupervised or supervised framework. In the following the state of the art feature extraction methods used in this thesis are explained.

4.2.1.1 Principal component analysis (PCA)

Principal component analysis is a classic method for unsupervised dimension reduction introduced in (Pearson, 1901). It is a sequence of transformations of a set of observations of possibly correlated variables into an orthogonal space where they are mutually uncorrelated and ordered in variance. The new transformed variables are called principal components (PC). Each PC is a linear combination of all original variables. In fact, the PCs are linear manifolds approximating the set of input points.

We consider a matrix of N observations with P variables $X_{N\times P} = \{x_1, ..., x_N\}$, with column-wise zero empirical mean. For a linear approximation of rank q , PCA can be used. Mathematically, the transformation is defined by a set of P-dimensional vectors of weights or loadings $v_k = (v_{k1}, \ldots, v_{kP}), k = 1, ..., q$ that map each row vector $x_i, i = 1, ..., N$ of X to a new vector of PC scores $t_i = (t_{i1}, \ldots, t_{iP})$, given by $\{t_{ki}\} = x_i \cdot v_k$. So that, the individual variables of t_i considered over the data set successively inherit the maximum possible variance from X, with each loading vector v_k constrained to be a unit vector.

The computation starts for the first loading and component as follows:

$$v_1 = \arg\max_{\|v\|=1} \left\{ \sum_{i=1}^{N} t_{ki}^2 \right\} = \arg\max_{\|v\|=1} \sum_{i=1}^{N} (x_i.v)^2 \tag{4.35}$$

writing this in matrix form we have:

$$v_1 = \arg\max_{\|v\|=1} \|Xv\|^2 = \arg\max_{\|v\|=1} v^T X^T X v \tag{4.36}$$

since v_k is a unit length vector, equivalently we can write:

$$v_1 = \arg\max_{\|v\|=1} \left\{ \frac{v^T X^T X v}{v^T v} \right\} \tag{4.37}$$

This is a Rayleigh quotient and for a symmetric matrix such as $X^T X$, the quotient's maximum possible value is the largest eigenvalue λ_1 of the matrix, which occurs when v_1 is the corresponding eigenvector.

With v_1 found, the first PC can be found in the transformed co-ordinate $t_1 = X.v_1$). Then, the kth component can be found by subtracting the first $k-1$

principal components from X:

$$\hat{X_K} = X - \sum_{S=1}^{K-1} X v_s v_s^T \tag{4.38}$$

and then finding the loading vector which extracts the maximum variance from this new data matrix:

$$v_k = \arg\max_{\|v\|=1} \left\{ \hat{X}_{k-1} v \right\}^2 = \arg\max_{\|v\|=1} \left\{ \frac{v^T \hat{X}_{k-1}^T \hat{X}_{K-1} v}{v^T v} \right\} \tag{4.39}$$

These calculations result in finding d number of loading or eigen vectors $V_{p\times d}$ and PCs $T_{p\times d} = XV$. The empirical covariance matrix $Q = X^T X = V \Lambda V^T$ which results in:

$$V^T Q V = \Lambda \tag{4.40}$$

where Λ is a diagonal matrix of eigen values $\{\lambda_1, ..., \lambda_d\}$.

Figure 4.6 shows an example of some 3D data projected to a 2D hyperplane that is the first two PCs surface.

4.2.1.2 Sparse PCA

In PCA, each PC is a linear combination of all P variables and the loadings are typically nonzero. This makes the interpretation difficult. Usually it is preferred to achieve both dimensionality reduction and variable selection together. One simple way to achieve this is to threshold the loadings so that the loadings with absolute values smaller than a threshold be set to zero (Cadima and Jolliffe, 1995). There are many research work for SPCA in literature. They are reviewed in the paper that is described in chapter 7 and appendix C. Two of them are briefly explained here.

An algorithm called SCoTLASS based on regression or reconstruction error property of PCs was developed in (Jolliffe et al., 2003). The procedure obtains sparse loadings by directly imposing an L_1 constraint on PCA. SCoTLASS successively maximizes the variance:

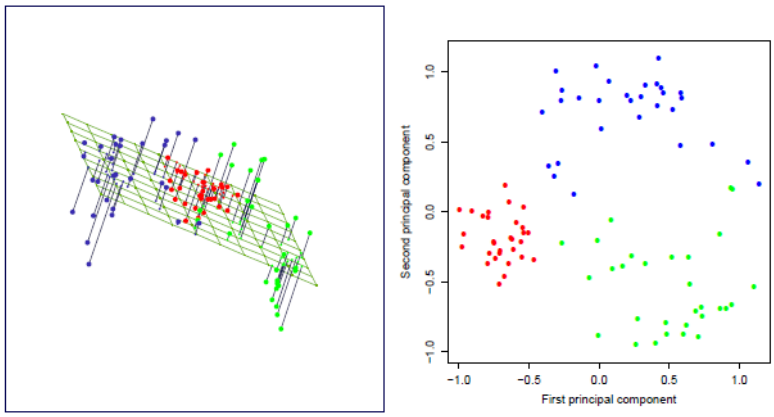


Figure 4.6: The best rank-two linear approximation of some 3D data (left). Illustration of the projected data onto the first two PCs surface (right) (Hastie et al., 2009).

$$v_k^T (X^T X) v_k$$

subject to:

$$\begin{cases} v_k^T v_k = 1 \\ v_h^T v_k = 0 \quad (for\ k > 2, h < k) \end{cases}$$

and the extra constraints:

$$\sum_{j=1}^{P} = |v_{kj}| \leq c \tag{4.41}$$

for some tuning parameter c. This last constraint can yield some exact zero loadings for sufficiently small c value.

In another work (Zou et al., 2004), an SPCA algorithm was proposed using the Elastic-Net framework for L_1 penalized regression on regular PCs using least angle regression (LARS). Considering the d first PCs, $T_{P \times d} = [t_1, ..., t_d]$ and

$B_{P\times d} = [\beta_1, .., , \beta_d]$, for any λ let:

$$(\hat{T}, \hat{B}) = \arg\min_{T,B} \sum_{i=1}^{N} \left\| x_i - TB^T x_i \right\|^2 + \lambda \sum_{j=1}^{d} \|\beta_j\|^2 + \sum_{j=1}^{d} \lambda_{1,j} \|\beta_j\|_1 \quad (4.42)$$

$$\text{subject to } T^T T = I_{d\times d} \quad (4.43)$$

Then $\hat{\beta}_j \propto v_j$ for $j = 1, ..., d$.

The same λ is used for all d components. However, $\lambda_{1,j}$ is different for penalizing the loadings of different PCs. If $P > N$, a positive λ is required in order to get exact PCA when the sparsity constraint (the lasso penalty) vanishes ($\lambda_{1,j} = 0$). For details about the solution, we refer to (Zou et al., 2004).

4.2.1.3 Supervised PCA

There are two different methods for supervised PCA (Bair et al., 2006; Barshan et al., 2011); In an earlier work (Bair et al., 2006), a pre-processing step was added to conventional PCA. So that, based on the regression coefficients of initial features, only a subset of features with higher scores are considered for PCA. In another work (Barshan et al., 2011) a generalization of PCA which aims at finding the PCs with maximum dependency to the response variables is proposed. In that work, the Hilbert–Schmidt independence criterion (HSIC) (Gretton et al., 2005) was used as the dependency function between the data and target response. It finds a sub-space XV such that, the dependency between the projected data XV and the target vector Y is maximized:

$$\max_V tr(KHLH) = \max_V tr(HXVV^T X^T HL) = \max_V (V^T X^T HLHXV) \quad (4.44)$$

where $H, K, L \in \mathbb{R}^{N\times N}, K_{ij} = k(x_i, x_j), L_{ij} = l(y_i, y_j)$ and $H_{ij} = I - N^{-1} ee^T$ is the centering matrix (e is a vector of all ones). Therefore, in order to maximize the dependency between two kernels, the value of the empirical estimate of HSIC, i.e., $tr(KHLH)$ is maximized. Thus, the following optimization problem

was solved in closed form using Eigen vector decomposition:

$$\arg\max_V tr(V^T X^T HLHXV) = \arg\max_V tr(V^T QV) \tag{4.45}$$

$$\text{s.t.} VV^T = I$$

If $Q = X^T HLHX$ is a symmetric and real matrix, with Eigen values $\lambda_1 \leq ... \leq \lambda_P$ and the corresponding Eigen vectors $v_1, ..., v_P$, then the maximum value of this cost function is $\lambda_P + \lambda_{P-1} + ... + \lambda_{P-d+1}$ and the optimal solution is $V = [v_P, v_{P-1}, ..., v_{P-d+1}]$. d is the dimension of the output space S. Then, the PCs are obtained as $T = XV$. More details about this method are explained in Appendix C.

4.2.1.4 Discrete Cosine transform (DCT)

DCT can be considered as a feature extraction strategy. It is an appropriate transformation in the field of signal processing. It was first introduced in (Ahmed et al., 1974) to be used in the image processing area for the purpose of feature extraction. It is widely used in image compression, audio and signal processing. DCT can be applied on signals of one or more dimension. The DCT transformation of a 2D signal $f(x, y)$ of size $N \times M$ is defined as follows:

$$C(u,v) = \frac{2}{\sqrt{NM}}\alpha(u)\alpha(v)\sum_{y=0}^{M-1}\sum_{x=0}^{N-1} f(x,y)\cos\left[\frac{\pi(2x+1)u}{2N}\right]\cos\left[\frac{\pi(2y+1)v}{2M}\right] \tag{4.46}$$

for $u = 0, 1, 2, ..., N-1$ and $v = 0, 1, 2, ..., M-1$ and $\alpha(i) = \begin{cases} \frac{1}{\sqrt{2}} & i = 0 \\ 1 & i \neq 0 \end{cases}$. It is clear that for $u = 0$ and $v = 0$, $C(u, v) = \sqrt{\frac{1}{NM}}\sum_{y=0}^{M-1}\sum_{x=0}^{N-1} f(x, y)$ which is the average value of the 2D signal that is called the DC value whereas, all other coefficients ($u, v \neq 0$) show the progressively increasing frequencies and are called the *cosine basis function*. These basis functions are orthogonal and independent, that is, none of the basis functions can be represented as a combination of other basis functions. Therefore, in the transformed matrix $C(u, v)$, non of the elements are correlated.

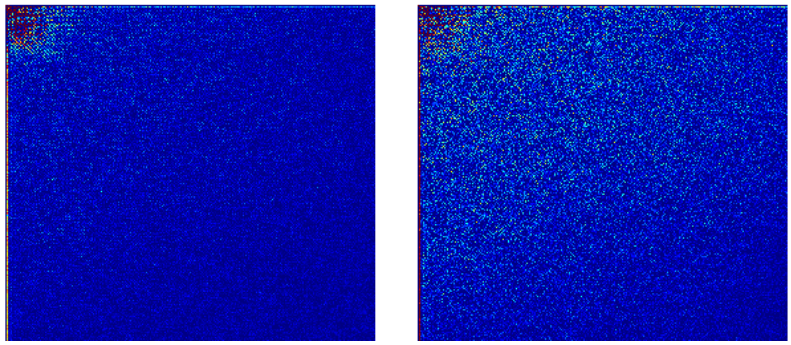


Figure 4.7: The 2D DCT map of the diffused reflectance images of milk products shown in figure 2.12

Besides its excellent decorrelation properties, DCT exhibits energy compaction for highly correlated data so that it can pack input data into as few coefficients. For example, the DCT transformation of the images shown in figure 2.12 is illustrated in figure 4.7. It shows how the images information are located mostly in a corner of their DCT map. In addition, it decomposes the spatial frequency in terms of various cosines transforms. In this thesis, DCT was employed to decompose the frequency information of the diffuse reflectance images of milk products described in section 2.2.4. That helped to characterize the products based on their high and low order DCT coefficients. The work will be explained in more detail in chapter 6 and appendix B.

4.2.2 Feature selection and testing

The main reason for feature selection is that the data contains many redundant or irrelevant features. This improves the interpretation and generalization of the model. A feature selection algorithm searches for a subset of features based on an evaluation measure which scores different feature subsets. Feature selection and feature extraction may seem similar but are different. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the original features. The selection algorithm varies depending on the evaluation metric, and the selection falls into one of the three main categories: wrappers, filters and embedded methods (Alelyani et al., 2013).

In wrapper methods, feature subsets are scored using a predictive model. Each new subset is used to train a model, which is tested on a hold-out set. Counting the number of mistakes made on that hold-out set (the error rate of the model)

gives the score for that subset. The wrapper methods train a new model for each subset, they are computationally intensive, but they usually select the best possible subset (Dy and Brodley, 2004).

In filter methods a proxy measure is used instead of the error rate to score the features (Alelyani et al., 2013). The measure is chosen so that, the computation cost do not be so high as it is in wrapper methods, whilst still capturing appropriate features. Examples of such a measure are mutual information, correlation coefficients, inter/intra class distance or the scores of significance tests for each class/feature combinations. Since the features are not selected based on the error of a specific type of predictive model, they are more general and may have lower prediction performance than a wrapper. However they are better for discovering the relation between the features.

In Embedded methods, feature selection is performed while a model is constructed such as lasso and EN explained in section 4.1.1.3 and 4.1.1.4 in which sparse linear models are constructed. In fact, the selected features are those with non-zero regression coefficients. The embedded methods tend to be between filters and wrappers in terms of computational complexity.

A feature selection can also be supervised or unsupervised. In this thesis, an unsupervised feature selection method is proposed for spectral data of food items that is described in chapter 8 and appendix D. In the following two important state of the art feature selection strategies are described.

4.2.2.1 Feature selection based on scale-space theory

Scale-space theory for signal analysis is a framework to find the local information of a signal (such as maxima and minima) when no prior information is available about it (Lindeberg, 1996). Therefore, the signal is represented at multiple scales to find the appropriate scales. In a multi-scale representation, structures at coarse scales constitute simplifications of corresponding structures at finer scales. In other words, the fine-scale information is successively suppressed or filtered as the scale increases. This principle preserves peaks or other feature to be artificially introduced through scales and forces the analysis to be from finer scale to coarser scales (Ceccarelli et al., 2009). Thus, the peaks can give information about the spectrum. The local extrema points are derived using a smoothing Gaussian kernel with varying scale parameter or standard deviation. Different strategies might be used for the choice of scale parameter, usually in a supervised framework, such as statistical tests (Tarn et al., 2008; Godtliebsen et al., 2002) or CV (Papandreou and Maragos, 2005; Ceccarelli et al., 2009). In this thesis, the scale-space method was tested in the technical report that is

described in chapter 8 and appendix D. In our work, a CV loop is used for the choice of scale parameter,

4.2.2.2 Feature subset selection using expectation-maximization (EM) clustering (FSSEM)

FSSEM is a wrapper method used for unsupervised feature selection. The idea is to cluster the data in each candidate feature subspace and select the best subspace with the minimum number of features (Dy and Brodley, 2004). This is done in three steps; feature search, clustering and feature subset selection criteria. In (Dy and Brodley, 2004) the sequential forward search (SFS) was used for feature search and the expectation maximization algorithm (EM) was used for clustering. Two different criteria were used for feature selection; Scatter separability criterion and maximum likelihood (ML). The number of clusters were found based on (Bouman, 1997) that begins a search for large number of clusters k_{max}, and then sequentially decrement this number by one until only one cluster remains. Among all pairs of clusters in step k, the two merged clusters are the ones that give the minimum difference in an objective function value. The objective function is based on the log-likelihood function with a penalty term added (see (Bouman, 1997)). For initialization of the EM algorithm, the sub-sampling initialization algorithm proposed in (Fayyad et al., 1998) was used. This algorithm is also tested in the technical report described in chapter 8 and appendix D.

4.2.2.3 Multiple hypothesis testing

Selecting features based on the scores of a significance test is a filter method as explained in previous section. Feature assessment based on multiple hypothesis testing is a statistical approach used for test and selection of features in problems that the number of features are very high compared to the number of observations $N \ll P$. It is mostly used for genomic data (Dudoit et al., 2003; Diz A. P., 2011) to assess the significance of individual features (genes). In this thesis, it is used for finding the significant features and the work is presented in chapter 10 and appendix F.

Considering to have M features and their p-value (e. g. by using the theoretical t-distribution probabilities, which assumes the features are normally distributed or a permutation distribution that does not make any assumption about their distribution), a hypothesis H is formed so that:

Table 4.1: Possible outcomes from M hypothesis tests.

	Called Not Significant	Called Significant	total
$H = 0$	U	V $(type - I)$	M_0
$H = 1$	T $(type - II)$	S	M_1
total	$M - R$	R	M

$$\begin{cases} H = 0 & \text{Negative(Null)} \\ H = 1 & \text{Positive} \end{cases}$$

This hypothesis is tested for all features $j = 1, ..., M$ and it is accepted $H_j = 1$ or in other words the result is significant at level α if $p_j < \alpha$. This test has $type - I$ error equal to α (for each individual test). That is, the probability of falsely rejecting $H_j = 0$ is α as shown in table 4.1.

Since there are a lot of individual tests (M is high), the overall measure of this error is quite high and should be corrected. One simple solution is the *Bonferroni* method. In order to reduce the number of false positive features (V), this method rejects $H = 0$ if the p-value of a feature satisfies $p_j < \frac{\alpha}{M}$. It is a useful method in cases that M is small, as it is based on the assumption that the co-variates are independent. However, in cases that M is quit high and high correlation exists between the co-variates, it is too conservative. That is, it calls too few features significant ($H = 1$). A more useful approach is the *Benjamin-Hachberg* (BH) (Benjamini and Hochberg, 1995) method. In this method the False Discovery Rate (FDR) is introduced as follows:

$$FDR = E(\frac{V}{R}) \tag{4.47}$$

It is the expected proportion of the false positive features V among the R features that are called significant. In this method, the FDR rate is bounded by a user defined level α. It is calculated based on the p-values obtained from an asymptotic approximation of the test statistic like a Gaussian or a permutation distribution.

If the hypotheses are independent, Benjamini and Hochberg showed that regardless of how many null hypotheses are true and regardless of the distribution of the p-values when the null hypothesis is false $H = 1$, this procedure has the

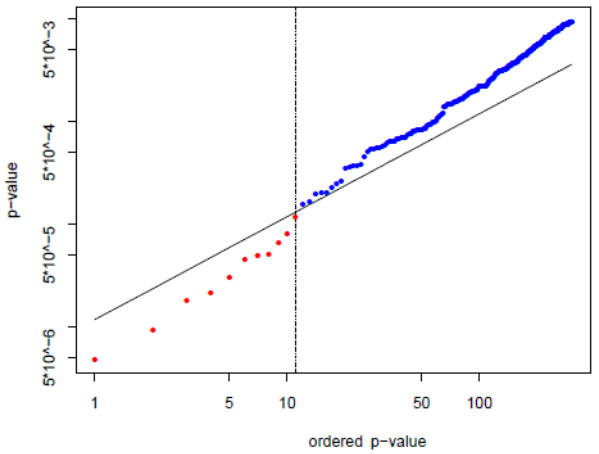


Figure 4.8: A plot of the ordered p-values $p_{(j)}$, the threshold line $(\alpha \frac{j}{M})$ as well as the critical point of the BH method (Hastie et al., 2009).

following property (Benjamini and Hochberg, 1995):

$$FDR \leq \frac{M_0}{M}\alpha = \alpha \tag{4.48}$$

In this method, the FDR is fixed at α level and the p-values are ordered $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(M)}$. Then a threshold point (L) is defined based on a threshold line $\alpha \frac{j}{M}$ $, j = 1, 2, ..., M$ so that:

$$L = \max \left\{ j : p_{(j)} < \alpha \frac{j}{M} \right\} \tag{4.49}$$

and the null hypotheses is rejected $H = 1$ for all tests that $p_j \leq p_{(L)}$, which is the BH rejection threshold (Hastie et al., 2009). As the FDR rate was kept fixed, the $type - I$ error is limited. This is illustrated in figure 4.8.

Multiple hypothesis testing was used in this thesis for finding the number of significantly changed features obtained from multispectral images of vegetables described in chapter 2.2.3. The work will be completely explained in chapter 10 appendix F

4.2.3 Over fitting

In training and learning a statistical model, over fitting can limit the generalization of performance and consequently affect the prediction capability on independent test data. Over fitting happens when the model fits well to the training data but has poor prediction ability on validation (unseen) data. This problem generally occurs when the complexity of the model is high, such as having too many parameters relative to the number of observations. To solve this problem, the performance of the model should be assessed during training to guide the choice of learning method or model and be sure about the quality of the ultimately chosen model. One of the widely used methods for this is cross validation (CV) that will be explained in the following.

4.2.4 Bias variance trade off

Considering a target variable Y , a vector of inputs X, and a prediction model $\hat{f}(X)$ that has been estimated using a training set T , the loss function for measuring errors between Y and $\hat{f}(X)$ is denoted by $L(Y, \widehat{f}(X))$ that can be defined as:

$$L(Y, \hat{f}(x)) = \begin{cases} (Y - \hat{f}(x))^2 & \text{squared error} \\ \left|Y - \hat{f}(x)\right| & \text{absolute error} \end{cases} \tag{4.50}$$

This loss function is calculated for both training and test sets where the test error or generalization error is the prediction error over an independent test sample. The expected prediction error (or expected test error) is the expectation averages over all test samples chosen randomly from the initial population. Figure 4.9 shows the average training and test errors using a lasso objective function. The data used in this example is the multispectral images of meat samples described in section 2.2.10 for prediction of a^* color component. As the model becomes more and more complex, it uses the training data more and is able to adapt to more complicated underlying structures. Hence there is a decrease in bias but an increase in variance. There is some intermediate model complexity that gives minimum expected test error. The training error is not a good estimate of the test error, as can be seen in figure 4.9, the training error consistently decreases with model complexity, typically dropping to zero if we increase the model complexity enough. However, a model with zero training error is over fitted to the training data and will typically generalize poorly.

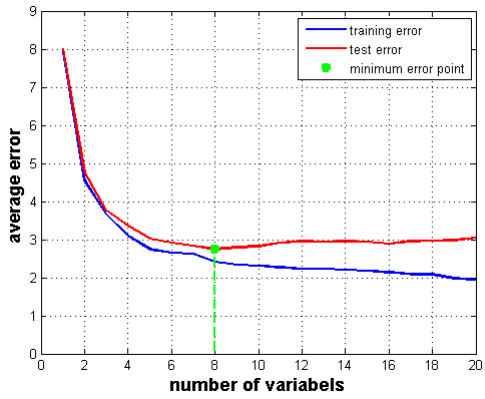


Figure 4.9: Behavior of the training and test error as a lasso model complexity increases.

Assuming that $Y = f(X) + \varepsilon$ where $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma_\varepsilon^2$, the expected prediction error can be written in terms of bias and variance:

$$\begin{aligned} E(X) &= E\left[Y - \hat{f}(X)\right]^2 = E\left[Y^2 + \hat{f}^2(X) - 2Y\hat{f}(X)\right] \\ &= E(Y^2) + E(\hat{f}^2(X)) - E(2Y\hat{f}(X)) \\ &= Var[Y^2] + E(Y)^2 + Var[\hat{f}(X)] + E(\hat{f}(X))^2 - 2f(X)E(\hat{f}(X)) \\ &= \sigma_\varepsilon^2 + Var[\hat{f}(X)] + (f - E(\hat{f}(X))^2 = \sigma_\varepsilon^2 + Var[\hat{f}(X)] + Bias(\hat{f}(X))^2 \end{aligned} \tag{4.51}$$

The first term σ_ε^2 is an irreducible error as it is the variance of the target around its true mean $f(X)$, and cannot be avoided no matter how well we estimate $f(X)$, unless $\sigma_\varepsilon^2 = 0$. The second term is the variance; the expected squared deviation of $\hat{f}(X)$ around its mean. The last term is the squared bias, the amount by which the average of the estimation differs from the true mean. Typically the more complex the model, the lower the (squared) bias but the higher the variance.

4.2.5 Model selection

Model selection methods are used to find the tuning parameter(s) of a model by estimating its expected test error. The tuning parameter varies the complexity of the model, and the average test error is used to find the best value that minimizes the error. After model selection by estimating the performance of different models to choose the best one, model assessment is performed to estimate its prediction error (generalization error) on new data.

One simple and most widely used method for estimation of the prediction error is the K-fold CV. The data is split into K roughly equal-sized parts. Then, for K times, one of the folds (parts) are kept as validation set and the model is fitted to the other $K-1$ parts of the data using the range of candidate values for the parameter(s). Then, the prediction error of the fitted model is calculated when predicting the kth part of the data. This is done for $k = 1, 2, ..., K$ and the K estimates of prediction error are averaged to find the best value for the unknown parameter based on the minimum error. Typical choices of K are 5 or 10. However, if the number of samples are limited, then the case $K = N$ is used, known as leave-one-out CV. In this case, $\kappa(i) = i$ and for the ith observation the fit is computed using all the data except the ith one. The choice of K depends on the size of data. A higher value for K reduces the bias of the CV estimator but increases the variance. In addition the computational cost is also high (Hastie et al., 2009).

In this thesis, the CV was used for training of the models. There are also other sampling and model selection strategies (Hastie et al., 2009) such as bootstrapping that in contrast to CV is a sampling approach with re-substitution and is useful when the samples diversity are high or Akaike information criterion (AIC) and Bayesian information criterion (BIC) that are based on a log-likelihood loss function (Hastie et al., 2009).

Chapter 5

Paper A - Supervised feature selection for linear and non-linear regression of $L^*a^*b^*$ color from multispectral images of meat

In food quality monitoring, color is an important indicator factor of quality. Supplying a consistent high quality product requires a continuous assessment in the meat industry. Conventional assessment methods in this case are based on subjective visual judgment and laboratory tests which are time-consuming, destructive and inconsistent in terms of human accuracy. In the case of meat, the most important quality criteria are visual appearances such as the texture pattern and the color of the meat. These parameters are linked to the chemical properties such as the water-holding capacity, intra-muscular (marbling) and protein content (Sun, 2010). As a result, surface color is an important parameter for quality measurement in the meat industry.

The CIELab ($L^*a^*b^*$) color space as a device independent color space is an appropriate means in this case due to its strong correlation with the human visual perception(Tkalčič and Tasič, 2003). The L^* is the luminance component and the a^* and b^* are the chromatic components. The commonly used colorimeter instruments can neither measure the $L^*a^*b^*$ color in a wide area over the target surface nor in a contact-less mode. However, developing algorithms for conversion of food items images into $L^*a^*b^*$ color space can solve both of these issues.

This paper addresses the problem of $L^*a^*b^*$ color prediction from multispectral images of different types of raw meat. The meat data for this work was provided by the Danish Meat Research Institute. Six different samples of meat from the used data set was shown in figure 2.8. Totally, we used 52 meat samples. The samples were divided randomly into training and test sets 25 times. In each data set, the number of training samples were 38. They were used for building the models and the remaining 14 samples were kept as unseen data for the test step. For each meat sample, multispectral images were acquired at 20 different wavelengths ranging from 430 to 970 nm using a VideometerLab. VideometerLab is a multispectral imaging device that was described in section 2.1.4. In addition, the reference measurements for L^*, a^* and b^* color components of each sample was available from Minolta measurements.

To form the feature vectors from the multispectral images, a Region of Interest (ROI) of size 200×200 pixel was selected from each sample image. In the next step, the pixel gray levels in each ROI were averaged at each wavelength. Therefore, we finally have 20 features per meat sample.

The efficiency of using multispectral images instead of the standard RGB is investigated. Furthermore, it is demonstrated that due to the fiber structure and transparency of raw meat, the prediction models built on the standard color patches do not work for raw meat test samples.

Three different regression strategies namely linear, nonlinear and kernel-based methods were used. Due to the limited number of samples, a five fold CV was applied on the training data for the optimal choice of model parameters in all of the methods. For linear regression, OLS, ridge, PLS, lasso and EN were used and for non-linear regression generalized feed-forward ANN with adaptive learning rate (CVHA), momentum BP (CVHM), Bayesian regularization (CVHB) and Neural regressor with quadratic cost function (CVHQ) were used. SVM was used as a kernel-based method.

Finding a solution that uses a minimum number of bands is of particular interest to make an industrial vision set-up simpler and cost effective. Therefore, besides the sparse regression methods such as lasso and EN, a supervised fea-

ture selection strategy is proposed that is combined by the regression strategies. The proposed method is based on the iterative use of lasso and EN. This feature selection method is compared with PCA as a pre-processing step.

The results showed that the proposed feature selection method outperforms the PCA for linear, non-linear and kernel-based methods. The highest performance was obtained by linear ridge regression applied on the selected features from the proposed Elastic net (EN)-based feature selection strategy. All the best models use a reduced number of wavelengths for each of the $L^*a^*b^*$ components.

The complete paper can be found in appendix A.

Chapter 6

Paper B - DCT-Based Characterization of Milk Products Using Diffuse Reflectance Images

In this paper, we proposed to use the two-dimensional Discrete Cosine Transform (DCT) for decomposition of diffuse reflectance images of milk products in different wavelengths. Two images of this kind have been shown in figure 2.12. These images were obtained by illumination of a hyperspectral coherent laser (460-1000 nm) into the surface of eight different milk products. They were milks and yogurts of different types and fat levels. This vision system was introduced recently for inspection of the structure of food items (Nielsen et al., 2011a,b). It is applicable for homogenous products where particle size and shape are important parameters. The main idea is to use the diffusion effects, which are known to be correlated to the microstructure. On the other hand, research findings in the field of food quality control have demonstrated a correlation between the texture, chemical and physical properties of food items with their microstructure characteristics (Bourne, 2002; Aguilera, 2005).

Considering these sequential relationships from the optical level to the quality

level, it is possible to build an automatic light-based system as a measuring tool, for monitoring the quality of dairies along the production line and avoid unwanted structures during the process. In addition, the use of a minimum number of bands is of special concern in this work. The reduction in the number of required wavelengths will assist to simplify the laser set-up and make the overall system simpler and cost effective.

There are two main visual effects in the hyperspectral images according to the characteristics of the milk products e.g. fat or viscosity. The main optical feature is the low frequency light diffusion emanating from the incident point that has the highest intensity in the image and another important effect is a high frequency speckle pattern caused by interference of coherent light due to surface irregularities (Goodman, 2007). Figure B.1 shows these two effects. These effects vary in different products according to their molecular composition and thus, reflectance and scattering properties of light.

In this paper, we propose to apply a DCT transform on the double logarithm of the entire diffuse reflectance image. DCT can decorrelate the highly correlated information in these images. It decomposes the low frequency diffusion effects and high frequency speckle effects into low and high order coefficients that could be quantified easier. The low order DCT coefficients are considered to characterize the optical properties. The entropy information of the high order coefficients are used to characterize the speckle effect.

In the next step, the discrimination power analysis (DPA) introduced in (Dabbaghchian et al., 2010), is employed as a selection criterion on the initial set of features for both wavelength and feature selection. It is a more careful method in terms of discrimination than the conventional zigzag or zonal masking for DCT coefficient selection. Especially, that is in our work, both the low and high order features are important. Using the final selected features of one proper wavelength, we could characterize and discriminate the eight different products. Comparing this result with the current characterization method based of a fitted log-log linear model, shows that the proposed method can discriminate milk from yogurt products better.

The complete paper can be found in appendix B.

Chapter 7

Paper C - Sparse supervised principal component analysis (SSPCA) for dimension reduction and variable selection

Principal component analysis (PCA) is one of the main un-supervised pre-processing methods for dimension reduction. Given a data matrix $X_{N\times P}$ with N data points and P features, it maps data into an orthogonal space based on the sorted variance of the input data. In the new space, each principal component (PC) is a linear combination of all original variables. The first principal component corresponds to the highest variance and the second to the second highest variance and so on.

In problems that the training labels are available, supervised PCA is a better solution. Because, PCA is unsupervised and although this is an advantage when the labels are unavailable, it can also be a limitation when a label or response vector is available. Because, it is not possible to guide the algorithm based on the target response. This is specially important when the task is regression or

classification, where it is preferred to map data into a space based on the data variations that depend on the response and not necessarily according to the maximum variation.

When both dimension reduction and variable selection are required, sparse PCA (SPCA) methods are preferred. This is the case, when the number of variables are very high and it is important to reduce the number of variables and remove any irrelevant or noisy variable. For example, in spectral imaging applications, each variable might be a wavelength and sparse PCs result in a simpler vision set-up or in biology, each variable might correspond to a specific gene and interpretation of the sparse PCs are easier. This also makes it possible to employ any suitable non-sparse data analysis method afterward. There are many research works for SPCA. They are reviewed in appendix C.

This work is focused on developing a sparse supervised PCA (SSPCA) algorithm. Such an algorithm will be appropriate for pre-processing of data sets for which a target response is available and a sparse solution for variable selection or interpretation is desired. The supervised PCA algorithm from (Barshan et al., 2011) is used to form an initial objective function. In order to find sparse solutions, penalization constraints for the Eigen vectors are considered. The resulting optimization problem is bi-convex and is solved using the PMD algorithm (Witten et al., 2009). Due to the use of a kernel in the objective function, the solution can handle data sets with linear as well as non-linear behavior. The sparse Eigen vectors found by the SSPCA algorithm can be used either for projection of a data set or feature selection. The projection is based on maximum dependency of the data to the target instead of its maximum variation. The proposed method for SSPCA is compared with PCA, the SPCA based on PMD algorithm and the supervised PCA method. The SSPCA objective function is close to the objective function of sparse partial least squares (SPLS) algorithm. Therefore, a comparison is also performed with SPLS. The experiments were conducted on both simulated and real data sets. The experimental results from the simulations as well as the real data sets demonstrate that the proposed algorithm for SSPCA can make an appropriate trade off between the accuracy and sparsity. It was almost best method in terms of sparsity compared to other methods and was better or comparable to the other methods in terms of accuracy.

The complete paper can be found in appendix C.

Chapter 8

Paper D - An unsupervised feature selection strategy for characterization of VIS-NIR spectral signals of food products based on local maxima

The spectral vision systems has found application in quality monitoring of food items widely. However, the spectra is usually obtained in high resolution and the spectral information are highly correlated. In addition, all of them are not relevant for the prediction task or may be noisy. Therefore, feature selection should be performed to exclude the irrelevant and redundant features and to reduce the complexity, dimensionality and over fitting problems.

In this report an unsupervised feature selection strategy is proposed based on the fact that, quality parameters of food items are related to their chemical composition or physical characteristics that influence their optical properties such as reflectance acquired by spectral measurements (Sun, 2009). Since the

dimensionality of the spectral features are high and they are highly correlated, feature selection is important for reducing their model complexity, improving their analysis result and interpreting the selected features. In this work, the significant local peaks in the spectrum that are related to the chemical or physical characteristics are used for prediction or classification of the quality parameters. Instead of all wavelengths only the local maxima are filtered and analyzed. As a result, the algorithm works faster compared to other selection methods that analyze all the features. In order to avoid small local fluctuations among the identified peaks, smoothing is performed prior to peak finding. This is important in cases where the spectra are noisy or the number of fluctuation on the envelope are considerable. This is performed based on adaptive thresholding of the wavelet coefficients of the spectra. Previously, a similar strategy was used for variable noise suppression of the spectral data (Schlenke et al., 2012).

The proposed strategy is compared to the state of the art scale-space strategy based on Gaussian filtering which is a supervised method and also utilizes the significant local peaks of the signal. We also compare our work to two unsupervised feature selection strategies ; a filter solution based on an entropy function and a hybrid solution as a combination of a filtering step based on feature clustering followed by a wrapper frame work that uses FSSEM (Feature Subset Selection using Expectation-Maximization (EM) clustering).

Three different data sets were used in this work for the experiments; spectroscopy measurements of apples that are used for prediction of their SSC content. They are described in section 2.2.2. The spectro-temporal features of milk fermentation process used for prediction of their fat level as described in 2.2.4 and the spectral data of the aquaculture feed pellets used for prediction of their astaxantin concentration level as described in 2.2.5.

The results show that the proposed method is superior than the two other methods in terms of accuracy and is better or comparable to the supervised scale-space feature selection method. In terms of computation time, the proposed method is considerably faster than all other methods.

The complete paper can be found in appendix D.

Part II

Application

Chapter 9

Paper E - A sampling approach for predicting the eating quality of apples using visible–near infrared spectroscopy

The use of visible and near infrared spectroscopy (VIS–NIR) for the rapid evaluation of fruit quality remains a topic of importance and interest for the food research community and food industry. It might be included in 'the tool box' for efficient farm management. Many research work have been performed on he use of (VIS-NIR) spectroscopy on quality prediction of fruits (see E).

Two of the most important fruit quality traits are SSC and acidity. These traits have a great influence on consumer liking and repetitive purchases. During fruit growth, the internal quality traits are expected to vary due to different causes (type of soil, weather, training and thinning techniques, etc.). This variation in quality might be the most important factor affecting the calibration models, which are used to train different spectroscopy devices. Model validation, an essential step to be carried out after calibration, has often been performed using

samples from the same batch. The use of a large sets of samples together with preprocessing statistical methods helps to obtain satisfactory results.

In addition, post-harvest sample arrangement is also important for the purpose of proper model construction. In this paper, a 'fractionator' tree sampling procedure is proposed to obtain representative apple fruit samples at time of harvest. These samples were used to evaluate the performance of VIS–NIR spectroscopy method for calibration and validation model development. Thus, the objectives of the study are: (1) evaluate the SSC and acidity prediction performance of an early and late season apple cultivar; and (2) to compare different sub-sampling techniques to form training and test sets on the overall performance of the prediction models. Furthermore, the main implications of the method in practice are discussed.

A total of 196 middle–early season and 219 late season apples 'Aroma' and 'Holsteiner Cox' samples were used to construct spectral models for SSC and acidity. PLS, ridge regression and EN models were used to build prediction models. Furthermore, we compared three sub-sample arrangements for forming training and test sets (smooth fractionator, by date of measurement after harvest and random). Using the smooth fractionator sampling method, combined with a supervised feature selection strategy that was proposed in the paper presented in appendix A followed by EN regression resulted in improved performance for SSC models of Aroma apples, with a coefficient of variation CVSSC=13%. The number of selected wavelengths by the feature selection method was 26. The model showed consistently low errors and bias:

$PLS/EN : R^2_{cal} = 0.60/0.60; SEC = 0.88/0.88 Brix; Bias_{cal} = 0.00/0.00;$
$R^2_{val} = 0.33/0.44; SEP = 1.14/1.03; Bias_{val} = 0.04/0.03$

However, prediction of acidity and SSC ($CV = 5\%$) of the late cultivar Holsteiner Cox produced inferior results as compared with Aroma.

Based on these results, it was possible to construct local SSC and acidity calibration models for early season apple cultivars. The overall model performance of these data sets also depend on the proper selection of training and test sets. The smooth fractionator protocol provided an objective method for obtaining training and test sets that capture the existing variability of the fruit samples for construction of VIS–NIR prediction models. The implication is that by using such efficient sampling methods for obtaining an initial sample of fruit that represents the variability of the population and for sub-sampling to form training and test sets, it should be possible to use relatively small sample sizes to develop spectral predictions of fruit quality. Using feature selection and EN appears to improve the SSC model performance in terms of R^2, $RMSECV$ and $RMSEP$ for Aroma apples. The complete paper can be found in appendix E.

Chapter 10

Paper F - Statistical quality assessment of pre-fried carrots using multispectral imaging

Multispectral imaging is increasingly being used for quality assessment of food items due to its non-invasive benefits. Since, there is a trend toward the use of these methods for quality control of food products such as meat, dairies and vegetables; it is of importance to test the capabilities as well as the reproducibility of such.

In this paper, the multispectral images of pre-processed carrots were used to detect the effect of storage on their color and NIR characteristics. The carrots were pre-fried without oil and then frozen for about two months. Then, they were moved to the refrigerator for experiments during a period of 14 days. Generally the surface color and texture are important parameters; indicating the quality of food. Multispectral images provide this information in visible bands and also more information about the subsurface and chemical characteristics in NIR bands. Using the multispectral images in visible and NIR bands, we tracked the quality of carrots during the storage days. The aim was to find out

in which days, significant changes occurred.

In this study, the preparation of carrots was performed in two steps. First, the vegetables were stir-fried (without oil) (Adler-Nissen, 2007). Research findings have shown that, stir frying produces high quality vegetables. After stir-frying, the products were frozen. Afterward, the pre-fried carrots were kept around two months in the freezer and then were moved into the refrigerator and their quality was evaluated within 14 days of storage. For quality assessment, the multispectral images of carrots were analyzed on days 2, 5, 8, 11 and 14.

Previously, the use of multispectral images for assessment of the color changes over time in per-fried vegetables was performed (Dissing et al., 2009). In this paper, high dimensional features were formed from the ratios of spectral bands and their corresponding percentiles. The high dimensional features based on band ratios are preferred, since they are more robust toward the undesired effects such as shadows. Multiple hypothesis testing was used to assess the high dimensional features. This method involves the significance assessment of the individual features. Since the dimensionality of the extracted features was quite high (3078), a conventional t-test at a significance level e.g. $\alpha = 0.05$ may find about 154 significant features just by chance even if, the null hypothesis of no change is true for all the features (Diz A. P., 2011). In our study, the False Discovery Rate (FDR) introduced in (Benjamini and Hochberg, 1995) and the expected number of significant features was used to detect the significant days of change. In addition, the SVM classification was employed. Although the classification results support the multiple hypotheses testing, it is difficult to use them alone, as a demonstration for significance of changes over the days. In addition, the method used in (Dissing et al., 2009) was applied to our data set, and the results were compared with the findings from the multiple hypothesis tests.

The experimental results show that the most important change in carrot samples occurred after 2 weeks. While with less significance level, they also changed after 2 days. Classification results obtained by SVM supported this. However, the EN regression results had high MSEs. As a result, the 2-sided t-tests on the regression predictions of any set of 2 days at a 5% level were significant.

In addition, considering the requirements of an industrial level vision system, it is interesting to know which wavelengths contributed most in the significant features. For this reason, we examined the frequency at which a wavelength was contributed into significant features. For example, for day 2 to 5, the three mostly used wavelengths were 435 nm (blue), 910 nm (NIR) and 470 nm (blue). In case of day 11 to 14, the 850-890 nm NIR bands as well as 660 nm (red) and 435 nm (blue) had the highest frequency. Similar analysis for other cases showed that, NIR bands as well as the blue and red wavelengths were among

the top frequent bands. The complete paper can be found in appendix F.

Chapter 11

Paper G - Optimal vision system design for characterization of apples using UV/VIS/NIR spectroscopy data

Quality monitoring of the food items by spectroscopy provides information in a large number of wavelengths including highly correlated and redundant information. The acquired optical characteristics such as reflectance or absorbance can represent the pigmentation and structural tissue changes in the plant organs.

There are different types of spectrophotometers used for spectroscopy and their spectral resolution (provided by monochromator) is an important characteristic showing the range of wavelengths they support (Bernd Herold, 2008),(Sun, 2010). However, not all the wavelengths are equally important for characterization of food items. Usually the data in adjacent wavelengths are highly correlated and many of them are redundant, whereas other wavelengths may not carry relevant information for the problem at hand. Therefore, choosing a proper set of wavelengths carrying relevant information will help to simplify the

vision system.

The aim of this paper is to solve such problems by employing sparse regression methods on UV/VIS/NIR spectroscopic data (306-1130 nm) of an apple cultivar. Two quality parameters, the sugar content called solvable solid content (SSC) and firmness of the apples were predicted using their spectroscopic data. Sparse regression methods assist to reduce the number of wavelengths (Hastie et al., 2009) and can simplify the vision set-ups used in food quality control. We compared three sparse regression techniques; least angle shrinkage and selection operator (lasso) (Hastie et al., 2009), elastic-net (EN) (Hastie et al., 2009) and fused Lasso (FL) (Tibshirani et al., 2005). The data set was divided into different training and test sets four times and the average results are considered. A 10-fold CV was employed for training the prediction models. However, using the model parameters corresponding to the minimum validation error resulted in the use of a considerable number of wavelengths. In order to reduce the number of wavelengths even more, two strategies were investigated in the training phase. First, the one standard error rule was used (Hastie et al., 2009). In addition, manual selection of the proper number of wavelengths corresponding to an acceptable performance compared to the optimal point was performed. It is shown that, considering a tradeoff between the number of selected bands and the corresponding validation performance during the training step can result in a significant reduction in the number of bands at a small price in the test performance. Both methods reduced the number of wavelengths significantly for all regression strategies. However, this reduction was more considerable for firmness than SSC. In addition, the second strategy decreased the number of required wavelengths more and achieved better performance than the first one.

Based on the results, the methods that are suitable for vision set-ups with different number of bands are determined. Besides the number of bands, the width of the regimes is important in spectrophotometer design. For example, lasso is suitable when a few individual narrow bands (less than 10 bands) can be provided by e.g. a few LEDs. EN is suitable when more bands (up to 200) in narrow regimes can be supplied. A monochromator capable of selecting a few narrow regimes of laser light suits this case. Finally, FL is the best choice when a lot of bands (e.g. more than 200) in broad regimes of laser light are available. The monochromator does not need to provide high resolution in this case. The complete paper can be found in appendix G.

Chapter 12

Paper H - Sensory quality prediction using multispectral imaging

The use of computer-vision based systems as non-destructive and in-line quality monitoring methods in food industry is increasing. The classic methods for food quality assessment are mainly based on laboratory tests and sensory evaluation, usually performed by human experts. However, such methods have some limitations. For example they can be destructive and they are dependent on well trained assessors.

Due to these limitations, the computer vision - based techniques such as multispectral imaging have been employed as an alternative for quality inspection of food items. These techniques are fast, non-invasive and result in reproductive quality monitoring methods in food industry. Additionally, they can be used objectively and in-line. Multispectral imaging gives information about the color and visual characteristics of the food under study as well as its chemical characteristics that are correlated to its quality (ElMasry and Sun, 2010). That is based on the unique spectral signatures of materials in the electromagnetic spectrum (Sun, 2009). Such spectral imaging systems can be designed very cheap for food quality monitoring.

In most cases the assessment is performed by detection or prediction of a "quality parameter" such as appearance condition (color or texture) or content level (sugar, acidity, etc.). Reviewing the literature shows that there are only a few research works on the use of vision-based systems for prediction of the human attitude about the food quality which we call "sensory attribute" or "sensory score".

Sensory analysis is one of the important methods for evaluation of the eating quality of food items and consumer satisfaction in food industry. Usually a panel of well-trained experts or untrained consumers evaluates a food product. There are several qualitative or quantitative sensory evaluation methods (Varela and Ares, 2012). However, sensory analysis in some cases is a destructive method and is time consuming. Therefore, it cannot be used as a routine analysis in an industrial production and processing line (Kamruzzaman et al., 2013).

This paper addresses the prediction of sensory attributes of wok-fried vegetables, (carrot and celeriac) using multispectral imaging techniques. Such kind of research for other types of food items were conducted before, that are reviewed in the paper presented in appendix H.

In this work, two types of vegetables (carrot and celeriac) were used for investigations. Two batches of stir-fried vegetables were evaluated after a freezing period followed by a chill-storage period for up to 14 days at 5 °C. At each day of experiment, the sensory evaluation was performed by a sensory panel of 6 assessors. In addition, multispectral images were acquired from the same samples in 19 different wavelengths (VIS-NIR).

The multispectral images were analyzed so that the vegetable pieces were segmented and high dimensional spectral features of 3078 length were formed per piece of vegetable. Using these features as input matrix X and the sensory attributes as the response vector Y, regression models formed for prediction of the sensory attributes and some strategies were employed to generalize the prediction models.

The results show that the sensory attributes that had some variation over the storage days and consistency over the two batches resulted in better models in terms of generalization. For carrot, the smell and for celeriac, the off-taste were the attributes that gave the best results. Based on this, the use of more batches and further samples can help to develop better prediction models in terms of generalization. In addition, analysis of wavelengths showed that, both visible as well as NIR bands were among the most contributing wavelengths in the image features that were used by the prediction models. However, the discoloration scores were not appropriate due to the limitation in human visual perception. Therefore, we conclude that a vision-based quality assessment system should

utilize multispectral images of some visible and NIR wavelengths together with an appropriate set of calibration sensory attributes (in this case excluding color), to improve the prediction task. In addition, the multispectral images provided a basis for assessing color changes not visible to the human eye. The complete paper can be found in appendix H.

Chapter 13

Conclusion

In this thesis several spectral datasets of different food items such as meat, diaries, fruits and vegetables were analyzed. In all cases the main challenge was to reduce the number of spectral wavelengths that directly influences the design of the required vision set-up. Depending on the type of the vision system employed, the spectral signals/images were available in tens to hundreds of wavelengths. In most cases, the number of available samples were smaller than the number of wavelengths ($N \ll P$) that forces an ill-posed problem. Different multivariate analysis techniques were employed to reduce the number of wavelengths. Using the reduced number of wavelengths, the spectral data was characterized for discrimination of different qualitative labels or prediction of a quantitative target.

Our strategy in this thesis was first to employ the existing linear and sparse linear methods (ridge, PLS, lasso, EN, FL, LDA) as well as non-linear and kernel based methods namely, ANN and SVM. Consequently, new feature selection and extraction algorithms were developed for the aim of wavelength reduction. A supervised feature selection method based on EN and lasso was developed. In addition, an unsupervised feature selection strategy based on local maxima of spectral 1D or 2D signals was proposed for the analysis of the spectral data of food products. For feature extraction, we proposed a novel sparse supervised PCA (SSPCA) method. Another feature extraction and wavelength selection method was introduced for characterization of the diffused reflectance images

based on DCT transform.

The supervised feature selection method combines the sparse solutions obtained by l_1 and l_2 regularizations of lasso and EN with an iterative strategy to discover the patterns of relevant informations from the frequent non-zero coefficients. The use of this method on two different datasets (meat and apples) demonstrated an improvement in prediction results.

The second proposed feature selection method utilizes the structure of the optical response profiles obtained along the electromagnetic spectrum and the fact that local maxima points are the best links to the phisio-chemical quality characteristics of material. This physical phenomenon was combined by signal processing and mathematical methods to develop a feature selection algorithm that is unsupervised and do not require any target response measurements. This method was also successfully tested on three different datasets of apples, diaries, and feed pellets of fish.

The SSPCA method was built based on maximization of an objective function with regularization constraint on the Eigen vectors. The resulting optimization problem is bi-convex and is solved iteratively based on soft thresholoding of Eigen vectors that produces sparse solution. This method was successfully applied on simulated and real datasets of food items and micro array.

The second proposed feature extraction and selection method was based on transformation of the diffused reflectance images into the DCT domain for decomposition of the high and low frequency feature effects in these kind of images. The final feature sets was formed in DCT domain using a few number of lower order DCT coefficients as well as the extracted entropy information. These features were used for both wavelength selection and characterization of the spectral images.

In addition, we have applied the statistical and mathematical modeling techniques on different scenarios of food related challenges; the effect of post harvest sampling, optimal spectrophotometer design, change detection in stir-fried vegetables over the storage time and sensory data prediction.

The method employed in this thesis were based on deterministic strategies while the effect of using probabilistic approaches remains for future investigations in this case. In addition, the spectral signature map of the food items that can be provided from chemometric measurements was not considered in this thesis. The proposed unsupervised feature selection method is closely related to this. The use of such spectral signature maps might improve this method.

Appendix A

Supervised feature selection for linear and non-linear regression of $L^*a^*b^*$ color from multispectral images of meat

Authors: Sara Sharifzadeh[1], Line H. Clemmensen[1], Claus Borggaard[2], Susanne Støier[2] and Bjarne K. Ersbøll[1].

1. Department of Applied Mathematics and Computer Science, Technical University of Denmark.
2. Danish Meat Research Institute, Roskilde, Denmark.

Published in *Engineering applications of artificial intelligence.*

Abstract

In food quality monitoring, color is an important indicator factor of quality. The CIELab ($L^*a^*b^*$) color space as a device independent color space, is an appropriate means in this case. The commonly used colorimeter instruments can neither measure the $L^*a^*b^*$ color in a wide area over the target surface nor in a contact-less mode. However, developing algorithms for conversion of food items images into $L^*a^*b^*$ color space can solve both of these issues. This paper addresses the problem of $L^*a^*b^*$ color prediction from multispectral images of different types of raw meat. The efficiency of using multispectral images instead of the standard RGB is investigated. In addition, it is demonstrated that due to the fiber structure and transparency of raw meat, the prediction models built on the standard color patches do not work for raw meat test samples. As a result, multispectral images of different types of meat samples (430-970 nm) were used for training and testing of the $L^*a^*b^*$ prediction models. Finding a sparse solution or the use of a minimum number of bands is of particular interest to make an industrial vision set-up simpler and cost effective. In this paper, a wide range of linear, non-linear, kernel-based regression and sparse regression methods are compared. In order to improve the prediction results of these models, we propose a supervised feature selection strategy which is compared with the Principal component analysis (PCA) as a pre-processing step. The results showed that the proposed feature selection method outperforms the PCA for both linear and non-linear methods. The highest performance was obtained by linear ridge regression applied on the selected features from the proposed Elastic net (EN) -based feature selection strategy. All the best models use a reduced number of wavelengths for each of the $L^*a^*b^*$ components.

Keywords: $L^*a^*b^*$ color space, Multispectral imaging, Regression, Sparse regression, Artificial neural networks, Support vector machine, Supervised feature selection

A.1 Introduction

Monitoring the quality of meat products is a significant concern in the food industry. Supplying a consistent high quality product requires a continuous assessment in the meat industry. This requires a development of on-line inspection methods for automation of the inspection process (Sharifzadeh et al., 2012b). Conventional assessment methods in this case are based on subjective visual judgment and laboratory tests which are time-consuming, destructive and inconsistent in terms of human accuracy.

The visual appearance; such as the texture pattern and the color of the meat are the main criteria for both the manufacturer and customer. These parameters are linked to the chemical properties such as the water-holding capacity, intramuscular (marbling) and protein content (Sun, 2010). As a result, surface color is an important parameter for quality measurement in the meat industry.

One efficient color space for quantification of food items is the CIELab or $L^*a^*b^*$ color space, due to its precise characteristics (Mendoza et al., 2006; Brewer et al., 2006). It is a device independent color space defined by the International Commission on Illumination - abbreviated as CIE in 1976. $L^*a^*b^*$ has a perceptually equal space. This means that the Euclidean distance between two colors in the CIELab color space is strongly correlated with the human visual perception (Tkalčič and Tasič, 2003). The L^* is the luminance component and the a^* and b^* are chromatic components.

Colorimeters and spectrophotometers are traditional instruments for measurements of colors such as $L^*a^*b^*$ in the food industry. They provide a quantitative measurement in a similar way to the human eye (Wu and Sun, 2013)(Balaban and Odabasi, 2006). Colorimeters, such as the Minolta chromameter or the Hunter Lab, are used to measure the color of primary radiation sources that emit light and secondary radiation sources that reflect or transmit external light (León et al., 2006). Therefore, color values are obtained optically but not mathematically. Before doing the measurements, the instrument is usually calibrated.

Traditional instrumental measurements can only measure the surface of a sample that is uniform and rather small (Balaban and Odabasi, 2006). Hence, they cannot completely represent the surface characteristics especially when it is non-uniform and highly textured as is the case for meat. In order to have a global representation of the target surface, computer vision techniques can be used to quantify the color (Wu and Sun, 2013). This leads to the formation of a 3D map of $L^*a^*b^*$ color values. Such a map represents the spatial characteristics of the whole surface instead of a small area. Color space conversion techniques can be employed to transfer an image into the $L^*a^*b^*$ space with the desired numerical and visual specifications. Thereby, the images of the meat samples from other color spaces such as RGB or CMYK can be transferred into $L^*a^*b^*$ space. In this way, it is possible to convert each image pixel into $L^*a^*b^*$ and therefore, generalize the representation.

Reviewing the literature shows that, conversion to $L^*a^*b^*$ was mainly performed using RGB images. In (Larrain et al., 2008; Mendoza et al., 2006) standard sequential transformation into XYZ color space and then from XYZ to $L^*a^*b^*$ was used for RGB images of beef and vegetables respectively. In (Fdhal et al., 2009), conversion for the RGB images of the standard color patches into $L^*a^*b^*$ was

performed using BPANN [1]. In (Cao and Jun, 2011, 2008), RBFNN [2] and GRNN [3] were used for conversion from CMYK color space to CIELab respectively.

The use of RGB images has some drawbacks. An RGB image, captured by a digital camera, is formed by filtering the incoming photons into three broad primary channels representing the color variables; Red, Green and Blue (RGB). These three variables are enough to describe a color sensation. However, the intensity recorded in each channel is an integration over a large range of wavelengths and therefore, two objects with different spectral radiant power distribution may seem to have similar colors in an RGB image. This is called metameric failure, which means matching colorimetrically under one illumination, but differ under another. It occurs when the spectral radiant power distribution of two objects are different, but the rough splitting of photons fails to observe this (Dissing et al., 2010). In addition, RGB is a device dependent color space and the color of an object may be slightly different in two different camera records.

Multispectral imaging is an alternative for solving these limitations. In a multispectral imaging system, the sampling frequency of the electromagnetic spectrum is high and images are formed in very narrow bands compared to the three broad intervals used in standard RGB imaging. Therefore, the distribution of incoming photons for each pixel is approximated correctly. Besides the visual bands that characterize the color information, the higher wavelengths such as NIR are related to the chemical characteristics. Therefore, spectral imaging has been widely used for food quality control applications (Gamal et al., 2009; Dissing et al., 2009; Sharifzadeh et al., 2013b).

So far, multispectral imaging has never been used in color conversion of food items. Color conversion using the spectral images can be done based on statistical predictive models. The advantage of such methods over the standard matrix transformation was investigated in (León et al., 2006). In that work, a sequential transformation was used for conversion of the RGB images of color samples into $L^*a^*b^*$. In addition, OLS [4] linear regression and ANN[5] with early stopping generalization were employed and their results showed that the ANN model obtained the best performance. In (Dissing et al., 2010), the multispectral images of the standard color patches were transformed into the CIE-XYZ using linear regression models.

This paper focuses on conversion of multispectral images (430-970 nm) of different types of raw meat into $L^*a^*b^*$ units. In the following, we explain the main

[1] Back Propagation artificial neural network
[2] Radial basis function neural network
[3] Generalized regularized neural network
[4] Ordinary least square
[5] Artificial neural networks

points investigated in this paper:

Since the food items can have variation, it is important to create and validate the prediction models on food products. Therefore, the use of real meat samples instead of the color patches for building the prediction models was investigated. Uncooked meat is translucent and transparent. Therefore the light reflected from it, not only comes from its surface but part of it comes from below the surface. Meat also has structure due to fibers with orientation. The color patches do not have structure and the light is reflected directly from the surface. Therefore, a model built on color patches do not work well on raw meat samples.

Due to the fact that the vision systems with their spectra are costly and not feasible to implement in the industry for online food productions, the sparsity is important and performing predictions using a minimum number of wavelengths would make the required vision system more cost efficient. Therefore, we propose a new supervised feature selection strategy based on EN and lasso [6] regression as a pre-processing step. The selected features were compared with PCA using three different regression strategies. A complete comparison between linear, non-linear and kernel-based regression methods was performed, which we did not see in the previous works. In order to have a general and fair judgment about the methods, the original data set was divided randomly into 25 training and test sets and the regression methods were tested on all of them and the average results were considered.

Finally, the results of the spectral images were compared with the RGB images.

The rest of the paper is organized as follows; section A.2 is about color description and section A.3 describes the data preparation. In section A.4, we describe linear, non-linear and kernel-based regression methods respectively. Section A.5 is about the proposed supervised linear feature selection algorithm. Experimental results are presented in section A.6. Finally, there is a conclusion for this paper in section A.7.

A.2 Color Description

In principle, there are two methods for describing color; The spectral and the tristimulus data description (X-Rite, 2004). Spectral data, describes the surface properties of the colored object. It demonstrates how the surface affects (reflects, absorbs, transmits, or emits) light. Conditions such as lighting changes, the

[6] Least angle shrinkage and selection operator

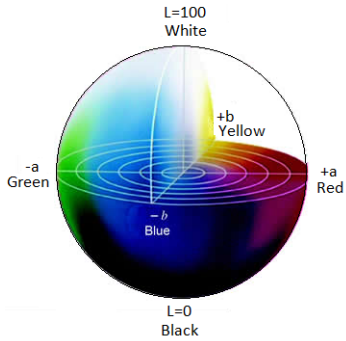


Figure A.1: $L^*a^*b^*$ 3D color space

uniqueness of each human viewer, and different rendering methods have no effect on these surface properties. In this paper, the multispectral images of meat are the input images.

The tristimulus data which is a 3D color space, describes the color of an object, as it appears to human eye or sensor, and as it would be reproduced on a device such as a monitor or printer. A CIELab color could be considered as a point in a 3D coordinate color space as shown in Figure 3.1. On the other hand, RGB and CMYK color representation describe a color as three values that can be mixed to generate the color. In contrast to these color spaces, CIELab is device-independent, meaning that the range of colors in this color space does not depend on the characteristics of a particular device, or the visual skills of a specific observer or the lightening condition. In addition, the RGB and CMYK color spaces are much smaller than the range of colors that is visible to the human eye.

In this paper, the output color is CIELab which is a uniform and widely used color scale. In this color space, L^* defines the lightness and ranges from 0 to 100; a^* denotes the red/green value; and b^* the yellow/blue value. The range of both chromatic components is between -128 and 128. This Color space resembles a three-dimensional space and uses rectangular coordinates based on the perpendicular yellow-blue, reen-red and illumination axes as shown in Figure A.1.

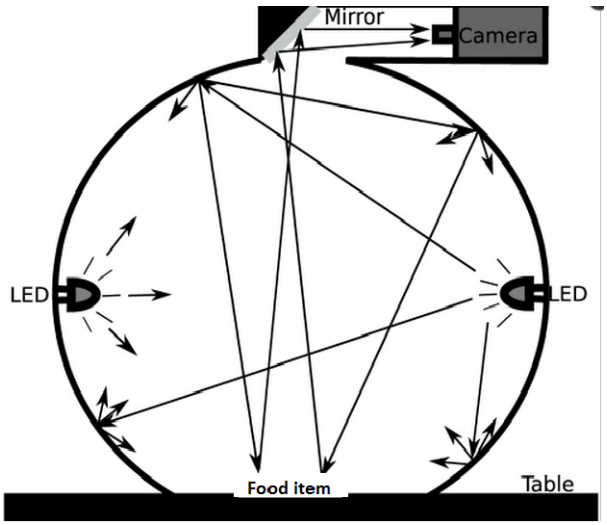


Figure A.2: Six different meat samples from the data set used in this paper

A.3 Data Preparation

The meat data for this work was provided by the Danish Meat Research Institute. Figure A.2 shows six different samples of meat from the used data set. In this data set, there were images of different types of turkey, chicken, beef, veal and pork.

In order to prepare the reference $L^*a^*b^*$ measure, two Minolta Chroma Meters CR300 and CR400 were used. Each Minolta data was acquired at 8 locations on each meat sample and the average and standard deviations of these readings were recorded. Then, the two Minolta results were averaged. The mean values were used as the reference L^*, a^* and b^* for each sample. The average standard deviations will be used as a reference for evaluation of the accuracy of the prediction models.

Totally, we used 52 meat samples which were divided randomly into training and test sets 25 times. In each data set, the number of training samples were 38 which were used for building the models and the remaining 14 samples were kept as unseen data for the test step.

For each meat sample, multispectral images were acquired at 20 different wavelengths ranging from 430 to 970 nm using a VideometerLab. VideometerLab is a

multispectral imaging device [7]. A sample is placed inside an integrating sphere. On top of the sphere, there is a camera which achieves a uniform and reproducible illumination. The illuminating diodes achieve the same level of intensity in all bands. They were calibrated radiometrically as well as geometrically to obtain the optimal dynamic range for each LED as well as to minimize distortions in the lens and thereby pixel-correspondence across the spectral bands. The optimal light condition avoids shadows and specular reflections (Dissing et al., 2009).

To form the feature vectors from the multispectral images, a Region of Interest (ROI) of size 200×200 pixel was selected from each sample image. In the next step, the pixel gray levels in each ROI were averaged at each wavelength. Therefore, we finally have 20 features per meat sample. The feature matrix is $X_{N\times P}$, where N denotes the number of samples and P is the number of wavelengths. The three output components are $L_{N\times 1}, A_{N\times 1}, B_{N\times 1}$. For ease of notation, we consider each of them as Y in the following sections.

One important point about the data set is that, we did not know the regions, where the measurements were performed. This means that, there is a deviation or mismatch between the regions from which, X is formed and the regions that Y values were measured.

In order to conduct the comparative experiment with the color checker, the standard X-Rite color checker was used. As shown in Figure A.3, it has 24 squares of colors in an 4×6 array. The multispectral images of this color checker were prepared in exactly the same wavelengths and light settings as the meat samples and the data set was formed in the same way. It has 24 samples in 20 wavelengths. The reference $L^*a^*b^*$ values of each color patch in the color checker is known.

A.4 Methods

In this section three regression strategies namely linear, non-linear and kernel-based methods that were used in this paper are explained. Due to the limited number of samples, a 5 fold CV was applied on the training data for the optimal choice of model parameters in all the methods.

[7] http://www.videometer.com

Figure A.3: The X-Rite color checker

A.4.1 Linear Regression

To convert the pixel intensities in the multispectral images into $L^*a^*b^*$ units, we can simply use the unbiased OLS model $\hat{Y} = X\hat{\beta}_{ols} + \varepsilon$, where ε is i.i.d. noise. However, since it is highly probable that some wavelengths have higher correlation to some of the output components ($L^*a^*b^*$), selection and shrinking strategies can be useful.

One simple regularization method is ridge regression which uses the L_2 norm penalty to shrink some of the regression coefficients. This decreases the variance of the outputs. Another efficient regularization method is PLS which selects directions or components based on both the variance in the co-variates and their correlation with the response (Hastie et al., 2009). If there is a lot of variation in X that has no connection to the variation of outputs and instead, the response is highly sensitive to the low variations of input, PLS can be a good solution. Therefore, we apply the PLS regression to improve the result, in the case such scenario exists in our data.

There are not necessarily prominent changes between images of all sequences of wavelengths and some of them are highly correlated. In this case a sparse solution such as lasso which uses the L_1 norm penalty can be employed:

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{P} |\beta_j| \right\} \tag{A.1}$$

Here, β_j is the j^{th} coefficient and λ controls the shrinkage rate. Another sparse regression method is EN. EN is in fact a compromise between lasso and ridge. Each regression coefficient is calculated as a weighted combination of ridge and lasso. EN selects variables like lasso, and shrinks together the coefficients of the correlated predictors like ridge (Hastie et al., 2009). The EN regression coefficients are computed by minimization of the following function:

$$\hat{\beta}_{EN} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda_1 \sum_{j=1}^{P} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \|\beta_j\| , \right\} \tag{A.2}$$

where there are both L_1 and L_2 penalty terms. The sparse regression methods result in the use of less wavelengths. As mentioned before, this is important regarding the economical concerns.

A.4.2 Non-Linear Regression

ANN can be used as a nonlinear regression solution. Figure A.4 shows the architecture of a simple ANN for regression with one hidden layer. First, M linear combinations of the input variables are built and then each combination is transformed using an activation function $h(.)$:

$$\phi_j(X) = h(\Sigma_{i=1}^{i=P} \alpha_{ij} x_k + \alpha_{0j}), j = 1, .., M \tag{A.3}$$

where α_{ij} is the weight parameter and α_{0j} is the bias. Then, the output $\hat{Y}$ is constructed as a linearly weighted combination of the nonlinear basis functions $\phi_j(X)$:

$$\hat{Y}(X;\beta) = f\left(\sum_{j=1}^{M} \beta_j \phi_j(X) + \beta_0 \right) \tag{A.4}$$

β_j and β_0 are the weight and bias parameters respectively, and $f(.)$ is an activation function which is usually, the identity function in the case of regression (Bishop, 2006).

Although this nonlinear model is more complex and difficult to interpret, it may probably be more accurate for some types of data. Therefore, when there is no

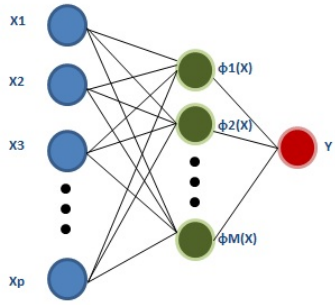


Figure A.4: The ANN diagram for regression with one hidden layer

need for a detailed interpretation of the model, ANN may be a good solution which is the case for color conversion. The choice of basis function and the solution strategy for the weight parameters vary in different ANNs. In addition, the architecture of an ANN is also based on the number of hidden layers and neurons. As mentioned in Section A.1, in many previous color conversion works, different types of ANN were used. Therefore, in this work, their application was investigated and compared with the linear methods.

A.4.2.1 ANN Modeling and Parametrization

One widely used ANN is the single hidden layer feed-forward ANN which uses a sigmoid basis function:

$$\phi_j(X) = \sigma_j(X) = \frac{1}{1 + \exp(-S_j X)} \tag{A.5}$$

Here, S_j is the scale parameter which controls the activation rate. A large scale may cause hard activation around 0.

Another type is the RBFNN that uses a non-linear RBF [8] based on the Euclidean distance or Mahalanobis distance (like a Gaussian kernel function):

$$\phi_j(X) = \rho_j\left(\|X - \mu_j\|\right) \tag{A.6}$$

[8] Radial basis function

Where μ_j is the center vector of the j^{th} hidden node and ρ is the distance function. The RBFNN also has one hidden layer.

The parameters of the ANN models are commonly estimated by minimization of the sum of square function shown in Equation A.7, using the BP procedure (Hastie et al., 2009). This is a gradient descent process.

$$E(\beta) = \min \sum_{n=1}^{N} \left\| \hat{Y}(X_n; \beta) - Y \right\|^2 \tag{A.7}$$

BPANN is a well known and widely used network and it has been used for color conversion problem as mentioned in Section A.1. Although it is a powerful algorithm, it has some drawbacks. One important problems with the error function minimization for complex and flexible models is the over-fitting on training data and poor generalization. Because a complex model is more flexible in capturing the training data behavior. Other problems are slow convergence and the possibility that the network converges to a local minimum. The ANN algorithms are also sensitive to the initial points and it is recommended to restart the algorithm several times for this reason. We applied the simple BPANN as well as the generalized BPANN with early stopping on our data set. They were also used in (León et al., 2006; Fdhal et al., 2009), for conversion from RGB into $L^*a^*b^*$ units. Although they worked fine in some of the 25 random sets, the results were poor for most of them and the average results were not satisfactory. This is because of the above mentioned problems. Due to this oscillating and unstable behavior of BP, we employed other types of BPANN.

In the literature, there are ANNs that employ different strategies to overcome these problems (Bishop, 2006, 2003; Hagan et al., 1996). In this paper we applied some of these strategies and compared their results; The ANN with Adaptive learning rate and momentum term was tested to accelerate the convergence. In addition, different regularized ANNs were used to constrain the parameters. In the following, the tested ANNs will be explained in detail.

A.4.2.2 ANN with Adaptive Learning Rate and Momentum Term

Considering the error minimization in Equation A.7, the gradient $\nabla E(\beta)$ can be obtained by means of back-propagation of errors through the layers. This gradient is used in the family of gradient training algorithms which iteratively

form:

$$\beta_{k+1} = \beta_k - \eta_k \nabla E(\beta^k), k = 0, 1, 2, ... \tag{A.8}$$

where β_k is the current weight, $-\eta_k$ is the learning rate and k is the step number and $-\eta_k \nabla E(\beta^k)$ shows the search direction. The BP gradient-based training algorithms minimize the error function using the above gradient decent or steepest descent method with constant, heuristically chosen, learning rate.

The learning rate determines how fast a network will learn the relationships between input and output patterns. A smaller value of the learning rate means a slower learning process. In fact, the optimal learning rate changes during the training process, as the algorithm moves across the performance surface. Therefore, the performance of the steepest descent algorithm would improve, if the learning rate changes during the training process. An adaptive learning rate attempts to keep the learning step size as large as possible while keeping learning stable (Hagan et al., 1996).

The idea about using a momentum BP is to stabilize the weight change and smooth the osculation in the trajectory. Therefore, a fraction of the previous weight change $\Delta\beta^k$ is considered in updating of the current weights β^{k+1}. Acting like a low-pass filter, momentum allows the network to ignore small local minima in the error surface and slide through them. It also speeds the convergence because, when all weight changes are in the same direction, the momentum amplifies the learning rate.

$$\Delta\beta^{k+1} = \gamma \Delta\beta^k - (1 - \gamma)\eta_k \nabla E(\beta^k), k = 0, 1, 2, ... \tag{A.9}$$

where γ is the momentum coefficient and should be between 0 and 1. This gives the system a certain amount of inertia since the weight vector will tend to continue moving in the same direction unless opposed by the gradient term.

Both the BP with adaptive learning rate and BP with momentum term were applied on the 25 data sets.

A.4.2.3 Regularization of ANN

Feed-Forward ANN Regularization The simplest regularizer is the quadratic in which, a penalty term is added to the error function and penalizes the sum

of weights toward zero similar to the regularization of the linear methods. The results of this method were acceptable on the validation sets and some of the test sets. However, the average test results were not satisfactory, showing very unstable and oscillating response on the different sets. This may happen due to the convergence in a local minimum.

These poor results will not be presented in this paper. Instead, the Bayesian regularization was used. It is an interesting approach which estimates the ANN parameters by a probabilistic approach (Bishop, 2006). Both the model output targets Y and parameters β are characterized as random variables with normal distributions. Then, the Bayesian rule is applied, to calculate their prior and posterior probabilities. Consequently, the predictive distribution of the output is obtained, using the sum and product rules for probabilities as shown in Equation A.10. For more details we refer to (Bishop, 2006, 2003).

$$P(\hat{Y} \mid X, Y_{tr}) = \int P(\hat{Y} \mid X, \beta).P(\beta \mid Y_{tr})d\beta \tag{A.10}$$

where, Y_{tr} denotes the data used for training the model. The averaging nature of the Bayesian method over many different possible solutions solves the over-fitting problem.

Another regularized ANN that was tested is the Nr_quadratic neural regressor with a quadratic cost function from DTU:toolbox (Kolenda et al., 2002a). This is a two layer feed-forward ANN with a hyperbolic tangent non-linear functions for the hidden layer and linear output layer. The weights of the ANN are optimized with a MAP[9] approach and the quadratic error function is augmented with a Gaussian prior over the weights. An adaptive regularization is used to prevent the over fitting. For more information, we refer to the documents provided in (Kolenda et al., 2002a).

BPANN are sensitive to the number of neurons in their hidden layers.

Too few neurons can lead to under fitting and too many neurons can cause over fitting. For this reason, for training of all the ANN algorithms described in Sections A.4.2.2 and A.4.2.3, loops are used for the best choice of the number of hidden nodes. Algorithm 1 shows the procedures used to train the ANN model. In each CV iteration, there is a loop on hidden nodes size. There is also another loop which repeats the training for each fold and each hidden node size several times. This will restart the network, training from different initial points and

[9] Maximum a posteriori

Algorithm 1 Training algorithm for ANNs described in A.4.2.2 and A.4.2.3

Inputs: Training data (X_{tr}, Y_{tr})
Initialization:

- HD=vector of hidden neuron size
- Rep=number of repetition times
- Initialize the 5 fold indices

Algorithm:
1. For cv=1,...,5 repeat:
   Divide the inputs into training and validation sets
2. For nhd=1,...,HD repeat:
3. For rp=1,...,Rep repeat:
   train the ANN with nhd number of hidden nodes
   calculate the training error matrix ($HD \times Rep$)

End loops 2 and 3
   Find the vector of minimum training error ($1 \times HD$)
   Find their corresponding trained ANN
   Use these ANN to calculate the validation error ($cv \times HD$)

End loop 1

- Find the minimum validation error
- Find the corresponding ANN with best nhd

Output: Best trained ANN and validation error

also helps to avoid falling in a local minimum. The output network from this algorithm will be used for the test data.

RBFNN Regularization For generalization of the RBFNN, the GRNN is used (Specht, 1991). In GRNN, the best prediction with minimum variance is obtained as the conditional mean value of Y_{tr} given X.

$$\hat{Y}(X) = E\langle Y_{tr} \mid X\rangle = \int_{-\infty}^{+\infty} Y_{tr} P(Y_{tr} \mid X) dY_{tr} \tag{A.11}$$

This could be calculated using the joint probability. GRNN uses a nonparametric approach to calculate the joint probability $P(X, Y_{tr})$ by a Gaussian isotropic kernel (Parzen window). The resulting probabilistic output is shown in Equation A.13. The numerator is the sum of the weighted training targets which contribute according to their joint probabilities with the input test sample, to form the output target. The denominator normalizes the solution.

$$\hat{Y}(X) = \frac{\int_{-\infty}^{+\infty} Y_{tr} P(X, Y_{tr}) dY_{tr}}{\int_{-\infty}^{+\infty} P(X, Y_{tr}) dY_{tr}} \tag{A.12}$$

$$\hat{Y}(X) = \frac{\sum_{i=1}^{N} Y_{tr}^i exp(-\frac{D_i^2}{2\sigma^2})}{\sum_{i=1}^{N} exp(-\frac{D_i^2}{2\sigma^2})} \tag{A.13}$$

where $D_i = (X - X_{tr}^i)^T (X - X_{tr}^i)$ and Y_{tr}^i, X_{tr}^i are the i^{th}training sample values. σ is the standard deviation of the Gaussian kernel and is called the smoothing parameter. As can be seen from this equation, the contribution weights are in fact the Mahalanobis distance of the test input from the training samples. This means that the closer training samples will contribute more in the prediction of the output target. The smoothing parameter has great effect on the output prediction. With larger σ, more training data will contribute in the target output than with a small σ. In each CV iteration, we loop over different σ values and repeated the training like in Algorithm 1, for the proper choice of σ.

A.4.3 Kernel-based Regression

SVM [10]was used as a kernel-based method for regression. SVM is characterized based on a maximum margin algorithm. Given the set of training data $\{(x_1, y_1), ..., (x_N, y_N)\}$, SVM finds a $f(x)$ function that has at most ε deviation from the actual target y. For this aim, the features are mapped to an M-dimensional feature space using non-linear basis functions ($h(x)$). Then, a linear model is constructed in this feature space:

$$f(x, \beta) = \sum_{m=1}^{M} \beta_m h_m(x) + \beta_0 \tag{A.14}$$

To estimate β_m and β_0, a new type of loss function called $\varepsilon - sensitive$ loss function is used:

$$V_\varepsilon(r) = \begin{cases} 0 & if\ |r| < \varepsilon \\ |r| - \varepsilon & \textbf{otherwise} \end{cases} \tag{A.15}$$

The objective function to be minimized is as follows:

$$\min_{\beta, \beta_0} L(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2} \sum \beta_m^2 \tag{A.16}$$

The second term in Equation A.16 controls the complexity level of the model. This optimization leads to a kernel based solution:

$$\hat{f}(x) = h(x)^T \hat{\beta} = \sum_{i=1}^{N} \alpha_i K(x, x_i), \hat{\alpha} = (HH^T + \lambda I)^{-1} Y \tag{A.17}$$

where $K(x, x_i) = \sum_{m=1}^{M} h_m(x) h_m(x_i)$. For more information, we refer to (Hastie et al., 2009).

[10]Support vector machine

A.5 The Proposed Supervised Linear Feature Selection

Feature selection can be used as a pre-processing step before all the explained methods. It helps to avoid over fitting by reducing the number of trainable parameters as much as possible.

Since the sparse linear regression methods perform both feature selection and regression together, it is not expected that a feature selection step improve their results. But, for non-sparse regression methods, it can be effective.

In the case of a feed-forward ANN, with a flexible number of hidden nodes, it is well known that the hidden layer can be regarded as taking the role of feature selection and dimension reduction. In in each CV iteration of Algorithm 1, the loop over the number of hidden nodes performs this selection properly. It has been demonstrated that CV is a successful model selection method (Shi and Xu, 2006). In addition, for ANN models, feature selection can be applied on the input variables $X_{N\times P}$ as a pre-processing step before the regression. It can be combined with any type of neural network.

One common dimension reduction method is PCA. It projects the variables orthogonally into a new space in which, they are sorted according to their variances. Therefore, it is possible to exclude features with low variance from the model. But, PCA is an unsupervised feature selection algorithm. This means that it does not consider the important information in the target values Y_{tr} and their dependencies to the training spectra X_{tr}. In addition, PCA is not a sparse feature reduction method. Because each principal component is a linear combination of all the variables.

However, according to the reasons described in section A.1, we are interested in using a minimum number of wavelengths. Although the sparse linear regression methods such as EN and lasso perform this, to improve the prediction results, we propose to use them for supervised linear feature selection. As described in section A.4.1, these methods will remove the redundant and irrelevant variables from the model, even with low or high variance. Algorithm 2 shows the different steps of our proposed supervised feature selection algorithm to form the reduced feature sets from EN.

In Algorithm 2, the vector $Freq$ was used to record the number of times each wavelength had non-zero regression coefficients and w was the vector of wavelengths. EN regression was repeated 4 times on each of the 25 input training sets. The 4 repetitions were done to cancel the effect of randomness in CV loops.

At the final iteration, the frequency of being non-zero for each of the 20 coefficients were obtained. The sorted $Freq$ vector shows the top frequent non-zero coefficients. Their corresponding wavelengths could be found in the re-ordered version of the w vector according to the sorted $Freq$. At this step, the number of wavelengths, to be used as the final selected features were determined. For this aim, another iteration over all possible candidate numbers (1 to 20) were tested.

Algorithm 2 The proposed algorithm for feature selection using EN

Inputs: 25 sets of (X_{tr}, X_{ts}, Y_{tr})
Initialization:

- *Freq*=vector of zeros (1×20)
- W=vector of the 20 wavelengths

Algorithm:
1. For all the 25 sets and for *rep=1,...,4* repeat:
    Compute β_{EN} by training an EN regression model with 5 fold cv
    Add one to the *Freq* elements with non-zero β_{EN}coefficients
End loop 1
    Sort *Freq* in descending order
    Re-order the corresponding elements in W with respect to *Freq*
2. For *i=1, ..., 20* repeat:
3. For all the 25 sets and for *rep=1,...,4* repeat:
    Compute the RMSE of regression on the training data using the corresponding first i wavelengths of W
End loops 2 and 3
    Average the RMSEs over the 25 sets and 4 iterations
    Find the index of the minimum average RMSE among the 20. (n)
    Select the first n top wavelengths from W, (sel_{EN}) and form the 25 sets of Xtr_{EN}, Xts_{EN}

Output: 25 sets of (Xtr_{EN}, Xts_{EN})

In the case of higher number of wavelengths (when $N << P$) this can be reduced to a limited candidate list. The average RMSE [11] was considered as a criterion for the final decision. The best number of features among the 20 candidates corresponds to the one with the minimum RMSE (n). Finally, the selected wavelengths were used to form the new training and test feature matrices. The same algorithm was used for feature selection by lasso. These two method were compared with PCA.

[11] Root mean square error

A.6 Experimental Results

In this section, first the evaluation criteria for prediction models will be introduced. Then, we will show the results from the experiments on the X-Rite color checker. In the next step, the results of applying linear, non-linear and kernel-based models on all the spectral data will be presented. Then, we will show the results of the same models on the selected features from both our proposed method and PCA. Since there were many tables of results, only the box plots are illustrated here and the complete tables are presented in the appendix. The RGB images experimental results will be shown next and also an $L^*a^*b^*$ image will be formed. Finally, there will be a discussion.

A.6.1 Evaluation Measures for Prediction Models

R-square (R^2), RMSE and ΔE measures are used for evaluation of the models.

R^2 is a statistical measure that shows the amount of data variation explained by a regression model. In order to calculate the R^2, RSS[12], TSS[13] and ESS[14] are defined as follows:

$$RSS = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2, TSS = \sum_{i=1}^{N}(y_i - \bar{Y})^2, ESS = \sum_{i=1}^{N}(\hat{y}_i - \bar{Y})^2 \qquad (A.18)$$

The most general definition of the (R^2) or coefficient of determination is:

$$R^2 = (1 - \frac{RSS}{TSS}) \times 100 \qquad (A.19)$$

In this definition, (R^2) is calculated based on the unexplained variance by the model or in other words the variance of the model's error.

RMSE shows the estimated standard deviation of the error and is calculated as follows:

[12] Residual sum of squares
[13] Total sum of squares
[14] Explained sum of squares

Table A.1: The average of the standard deviations over the 25 test sets for L*a*b* components

	The average standard deviation
L*	2.237
a*	1.115
b*	0.879

$$RMSE = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{A.20}$$

As mentioned in Section A.3, the average standard deviation of the Minolta measurements can be used as a reference for evaluation of the prediction models. Table A.1 show the overall average of standard deviations for all the 14 samples in the 25 test sets. The estimated RMSE as the standard deviation of the prediction model, can be compared with these measured values.

The delta error ΔE shows the color difference. A ΔE of 1 or less is not perceptible by human eye. A ΔE between 3 and 6 is typically considered as an acceptable match in commercial applications. Since the ΔE calculations are illuminant-dependent, calculations from colors viewed or measured under different illuminants are not comparable (Upton, 2006).

$$\Delta E = \sqrt{(L - \hat{L})^2 + (a - \hat{a})^2 + (b - \hat{b})^2} \tag{A.21}$$

A.6.2 Color Checker Test Results

As described in Section A.1, due to the transparency and texture structure of the raw meat, the use of multispectral images of meat may probably work better than the standard color checkers for color prediction. This was investigated by performing two experiments on the color checker data and meat samples.

In the first experiment, the color checker data was used for training a prediction model and in the second one, the 25 meat training sets were used. Then, they were applied for prediction on the respecting training sets as well as the 25 test sets. The average results were considered. The linear sparse EN regression was used to form the prediction model for $L^*a^*b^*$ color components. Since the color

Table A.2: The training and test results of the prediction model built on the color checker and meat data

LOOCV-EN	Color Checker Model			Meat Model		
	L*	a*	b*	L*	a*	b*
$R^2_{tr}\%$	93.25	95.55	95.71	90.73	94.85	83.92
$RMSE_{tr}$	4.75	5.08	6.89	2.50	1.25	1.02
$R^2_{ts}\%$	84.06	-482.42	-521.64	87.63	87.28	68.07
$RMSE_{ts}$	3.18	12.20	6.26	2.78	1.76	1.41
ΔE_{ts}	12.35			3.22		

checker data had limited number of samples ($X_{24\times20}$), LOOCV [15] was used for both experiments on the color checker and meat data. This helps to have good generalization while finding the optimal model parameters. The results are presented in Table A.2.

As can be seen, both models were capable of predicting on their own training data. But, the color checker failed to predict the color components for the meat data as expected regarding the physical characteristics of the raw meat. The negative R^2 shows the high RSS in Equation A.19. These results motivate us to use the multispectral images of meat to build the prediction models.

The errors in the case of the color checker training data was higher than expected. The reason for this was investigated by calculation of 0.95% confidence interval of the mean values of the color patches. First, the standard error of the regions of interests in the 24 patches of color was calculated from which, we computed $\triangle x_{24\times20}$ for the 95% confidence interval of the mean values. Then, it was used for calculation of the confidence intervals for the three components and the averaged results were considered.

$$\triangle L_{24\times20} = \triangle x_{24\times20}\beta_{L(20\times1)} \rightarrow \overline{\triangle L} = 2.67 \tag{A.22}$$

$$\triangle a_{24\times20} = \triangle x_{24\times20}\beta_{a(20\times1)} \rightarrow \overline{\triangle a} = 7.34 \tag{A.23}$$

$$\triangle b_{24\times1} = \triangle x_{24\times20}\beta_{b(20\times1)} \rightarrow \overline{\triangle b} = 7.90 \tag{A.24}$$

These results explain the reason for high $RMSE_{tr}$ for the color checker. In addition, The average values of the ROIs for the 24 color patches plus/minus

[15] Leave one out cross validation

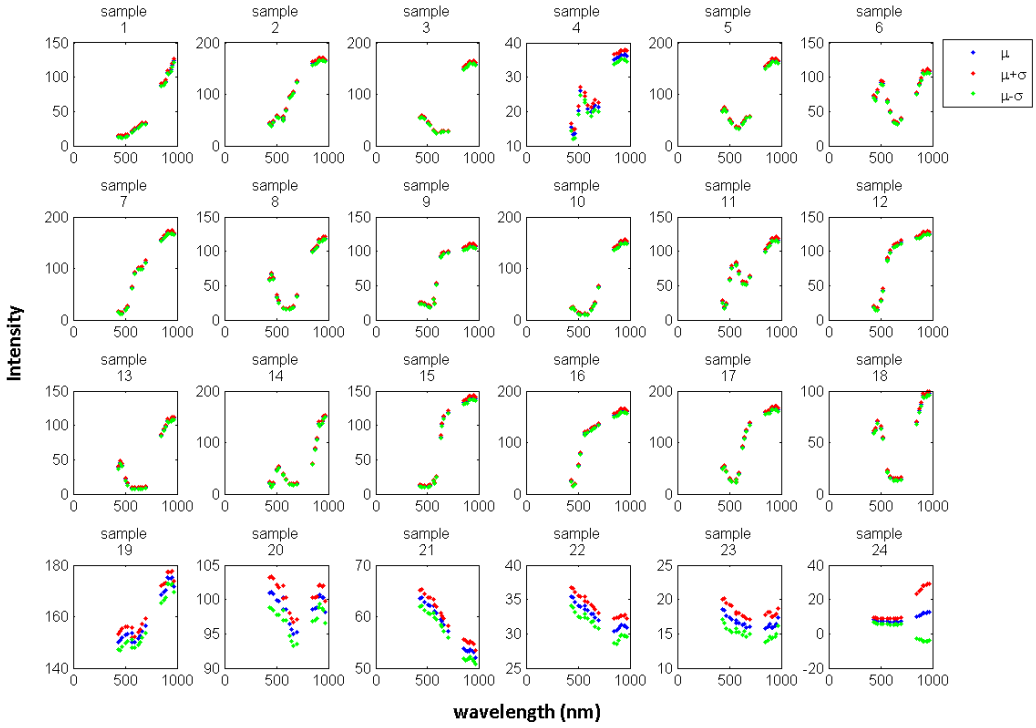


Figure A.5: The plot of the $(\mu \pm \sigma)$ in 20 wavelengths for the 24 ROI of the color patches. The horizontal axis shows the wavelength.

the standard deviation within each region, along different wavelengths $(\mu \pm \sigma)$, are plotted in Figure A.5. It shows that, although the color patches seem to be uniform, there is still variation in the spectral images of each color patch.

A.6.3 Linear Model Results

In this section the results of applying the linear regression methods described in Section A.4.1 are presented. As stated before, the tables of average results on the 25 training and test sets are shown in the appendix. In Figure A.6, the box plots of the R^2 of the test results over the 25 different sets are shown. The R^2 results for the L^* and a^* components were better than b^* component. The test RMSEs (see the appendix), show higher prediction error compared to the measurements errors shown in Table A.1. The training set results was better than the test set. The best ΔE_{ts} was 3.12 obtained from the ridge regression.

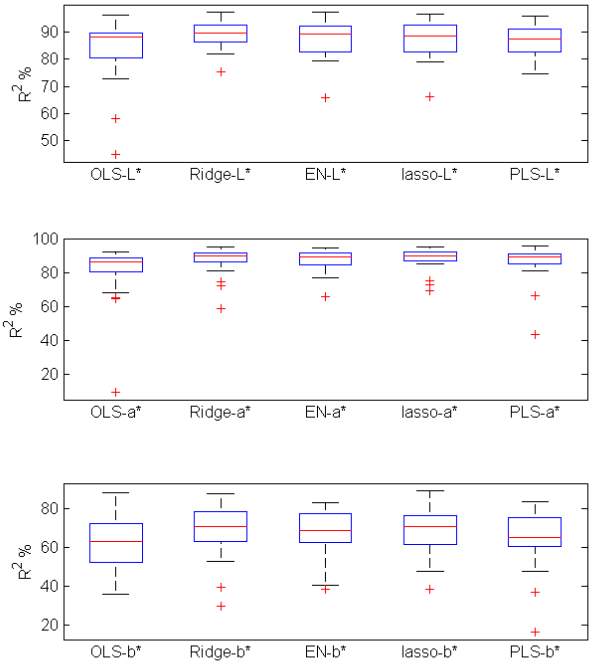


Figure A.6: R^2 box plots of the $L^*a^*b^*$ prediction for liner models on the 25 random test sets

Since the 25 sets were generated randomly, possibly some of the training sets did not include the existing variation inside the original data set. Considering the fact that the original data set consists of a few samples of different types of meat, the above mentioned issue, may explain some far data points from the median in the box plots.

Since we are interested in sparse solutions, the number of times that the EN and lasso regression coefficients were non-zero in the 25 sets are illustrated for the three components in Figure A.7. We call this a frequency map because, it shows the frequency of having non-zero coefficients for each wavelength. Comparing the wavelengths with the spectrum of colors shown in the bottom of the plots, helps to find which wavelengths are mostly selected by EN and lasso. As can be seen, some near infra-red wavelengths in all cases were among the top most frequent bands.

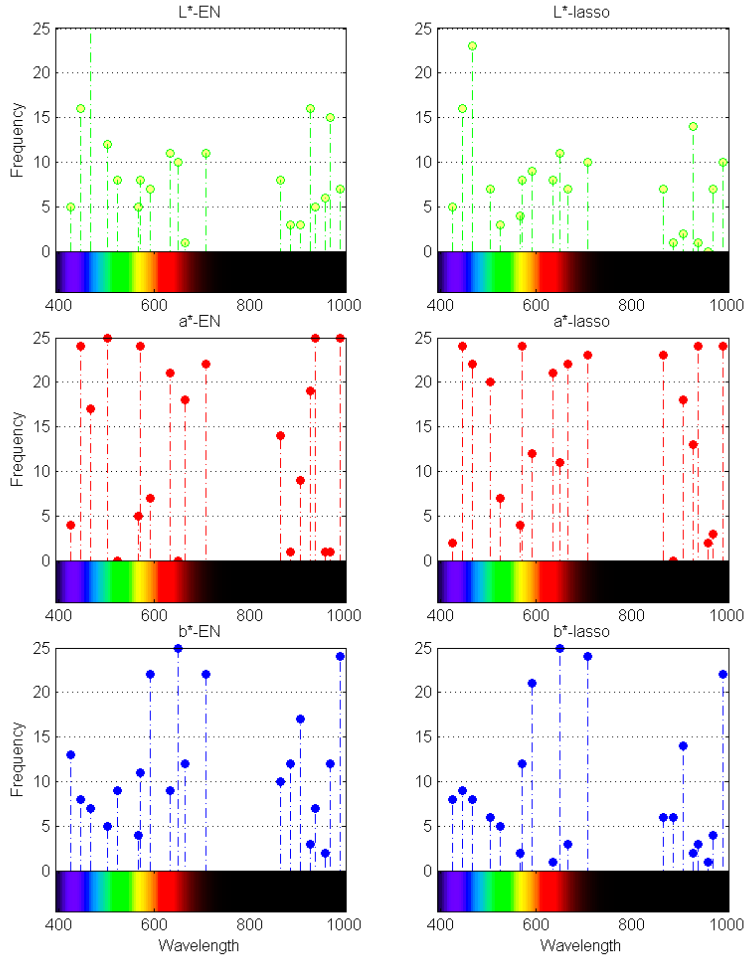


Figure A.7: The frequency map of the selected wavelengths by EN (left) and lasso (right)

A.6.4 ANN Results

In this section, the results of applying the non-linear regression methods described in Section A.4.2 are presented. Since in this paper different ANNs are compared, their names are contracted for the ease of notation. For feed-forward ANN, a simple one hidden layer architecture similar to the Figure A.4 was considered. The algorithm shown in Algorithm 1 was used for training the generalized feed-forward ANN with adaptive learning rate (CVHA), momentum BP (CVHM), Bayesian regularization (CVHB) and Neural regressor with quadratic cost function (CVHQ). The range of hidden neurons sizes were $\{5, 10, 20, 40, 60, 80, 100\}$. Similar algorithm was used for training the GRNN (CVSG). However, a loop for the best choice of the smoothing value σ was used instead of the hidden neurons loop. The regularized RBFNN model is a 2 layer network. For the smoothing value σ, 100 different values were generated logarithmically between 0.01 to 10.

Figure A.8 shows the box plots of R^2 test results. We can see that, there are some very far outliers from the median which may affect the overall average results significantly. Such a case can be seen for example, for the CVHB prediction for b* component. This may happen in ANN due to the inappropriate initial point or a convergence to a local minimum. Among the tested ANNs, the GRNN (CVSG) shows the lowest performance. Like linear models, the non-linear models work fine on the training data. The best training results are obtained from the CVHQ, CVHB and CVHA and for the test data, the best two models are the CVHQ and CVHB. The best ΔE_{ts} was 3.85 obtained by CVHQ. The average training results are satisfactory however, the test results are not better than the linear models using all the 20 wavelengths (see the appendix). One reason can be the high number of input variables. Regarding the higher complexity of the ANNs than the linear models, reducing their input variables may improve the results.

A.6.5 SVM Results

Figure A.9 shows the box plots of R^2 test results for the three components using SVM. The results of the SVM regression model does not show a significant improvement compared to the previous methods. During training the model, a linear kernel obtained the best result and was used in the final model. In contrast to the previous models, there are no outliers in the output results.

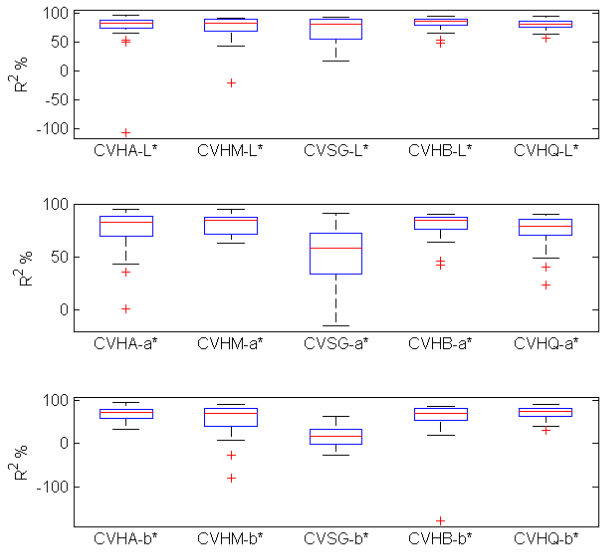


Figure A.8: R^2 box plots of the L*a *b * prediction for non-linear models on the 25 random test sets

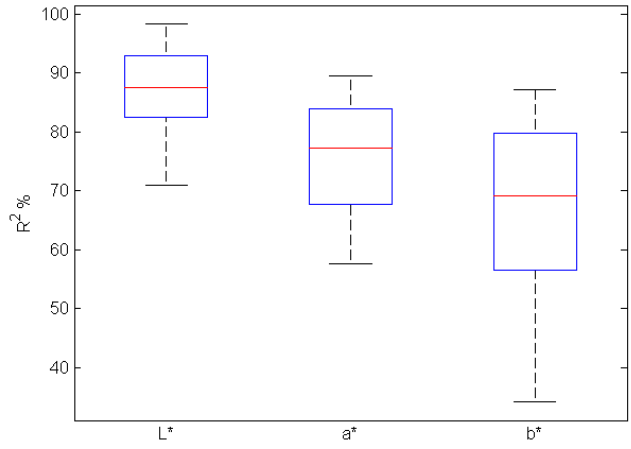


Figure A.9: R^2 box plots of the L*a *b * components from SVM prediction results on the 25 random test sets

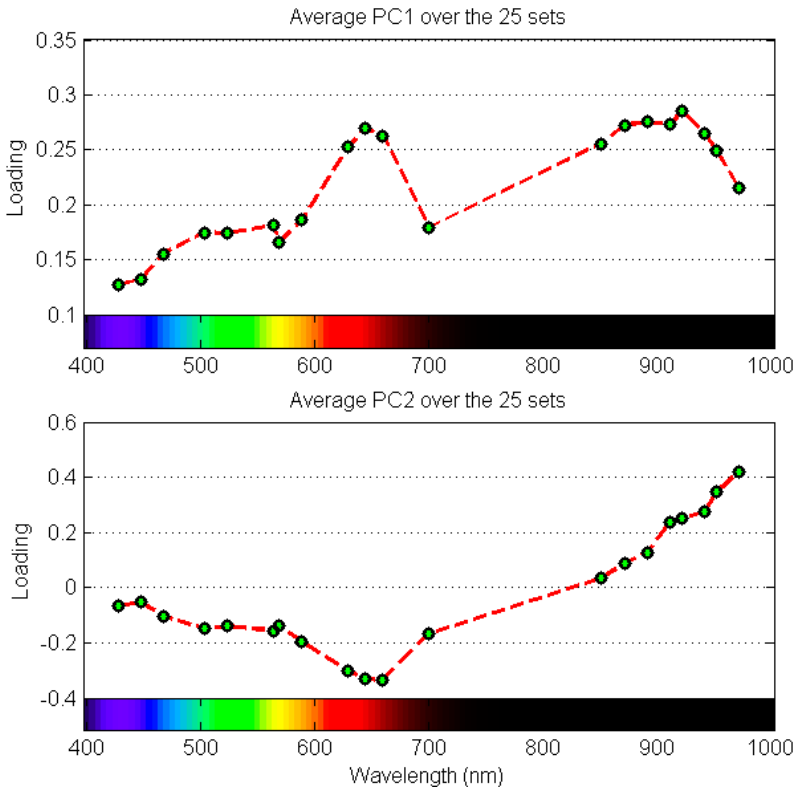


Figure A.10: The average of the first 2 PCs from the 25 random data sets versus the wavelengths

A.6.6 Feature Selection Results

The proposed supervised feature selection strategy based on EN and lasso in Section A.5 as well as PCA were used to reduce the number of wavelengths. Then the resulting reduced spectral data was employed in training the models.

First, a PCA analysis was performed on each of the 25 data sets. The 97% of the variation was explained just by the first two PC components in all cases, which was a very significant reduction in data dimension. Figure A.10 shows the average of the selected PCs in the 25 data sets with respect to the wavelengths.As can be seen, both PCs enhance the higher part of the wavelengths corresponding to the NIR wavelengths. The first PC which describes more than 90% of the

Table A.3: The number of top wavelengths selected by EN and lasso

	EN	lasso
L*	16	12
a*	8	12
b*	13	13

variations has another peak around the red color area, that corresponds to the different color ranges of the meat samples and can explain the correlation with the a^* component. However, the second component shows a negative correlation peak in the red color area. It also has two small peaks in blue and yellow ranges which explains the b^*color component.

The second and third sets of reduced features were formed using the Algorithm 2 in a supervised approach. Figure A.11 shows the frequency map of the 20 wavelengths by EN and lasso. Similar to the PC components, the near infra-red wavelengths have high frequencies in all cases specially, for the a^* component. In addition, some visible bands were among the high frequent wavelengths. These frequencies were sorted in a descending order and their corresponding 20 wavelengths were also re-ordered. Then, for each of the 25 training sets, a candidate subset of the top wavelengths were considered and an EN regression was applied for 4 iterations. The candidate subset length was varied from 1 to 20. The average RMSE results of these 20 candidate subsets are illustrated in Figure A.12. The minimum RMSE corresponds to the best number of top wavelengths. Table A.3, shows the final number of selected bands for each component.

The reduced sets of features obtained from the PCA and the proposed method were used to build the prediction models. Figure A.13 shows the box plots of the R^2 test results for linear, non-linear and SVM regression methods. This figure just shows the results of the EN-based feature selection. In the case of linear models, we can see that by using less bands, the results are better than Figure A.6, except for the two sparse methods, EN and lasso, as we expected. Comparison of Figure A.8 with this figure shows that, the use of less wavelengths did not made considerable changes in the median for non-linear models. Many outliers can be seen in the both box plots of the ANN methods. The lowest median among all methods was for CVSG in all the three components. Comparing Figure A.9 for SVM with this figure does not show important differences.

The complete results are presented in the appendix. The PCA did not improve the results in almost all cases. Comparing the results for the ANN models show some improvements in the maximum averages obtained on the test sets. This does not mean that all the non-linear models results were improved by the

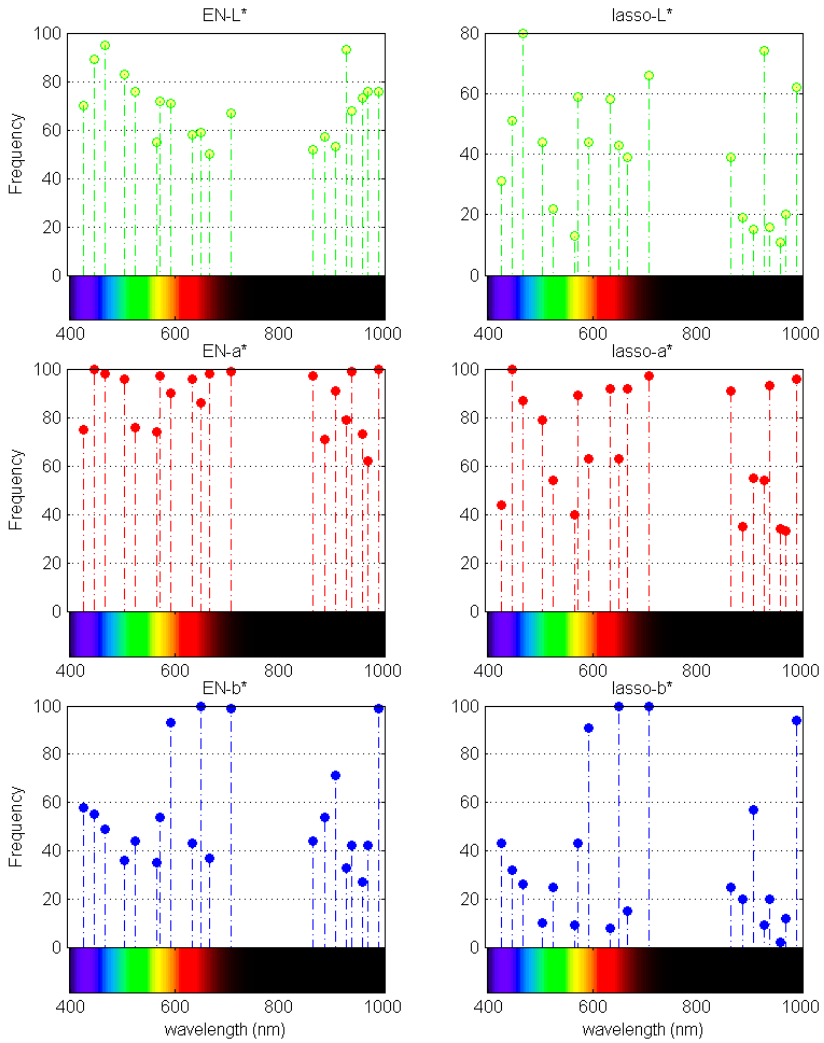


Figure A.11: The frequency map of the selected wavelengths by EN and lasso in 4 iterations for each of the 25 sets

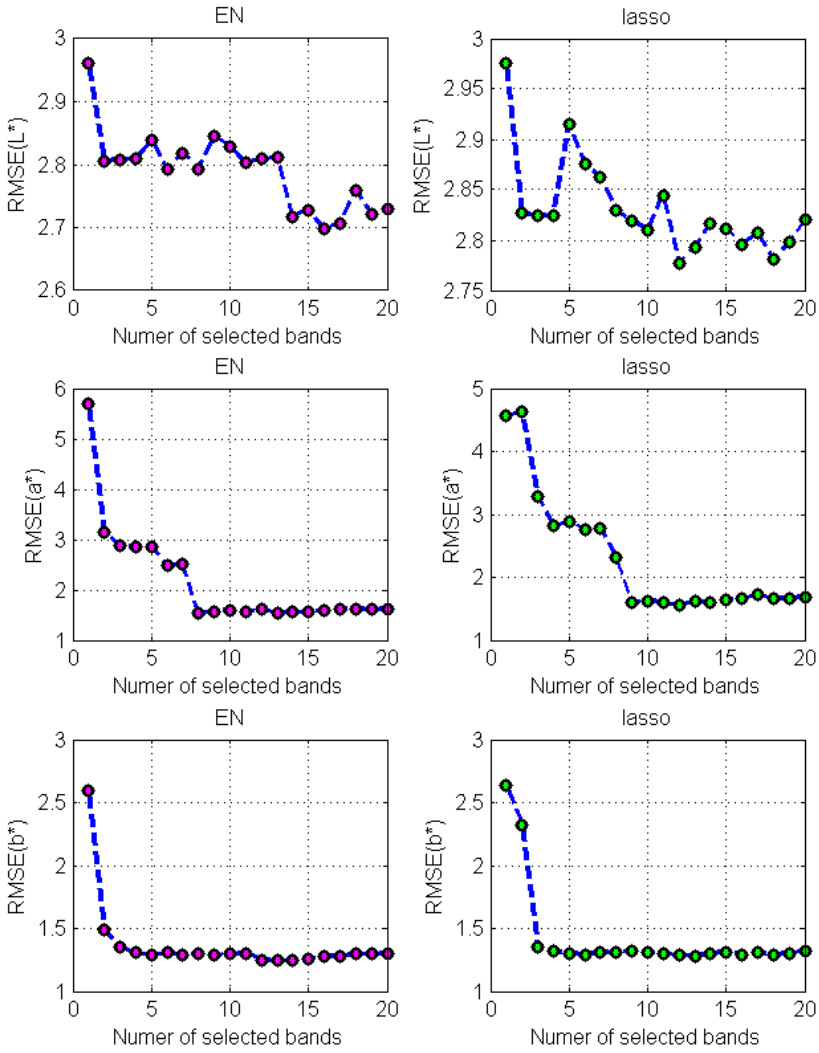


Figure A.12: The average RMSE results of EN and lasso regression for 20 candidate subsets of the sorted wavelengths

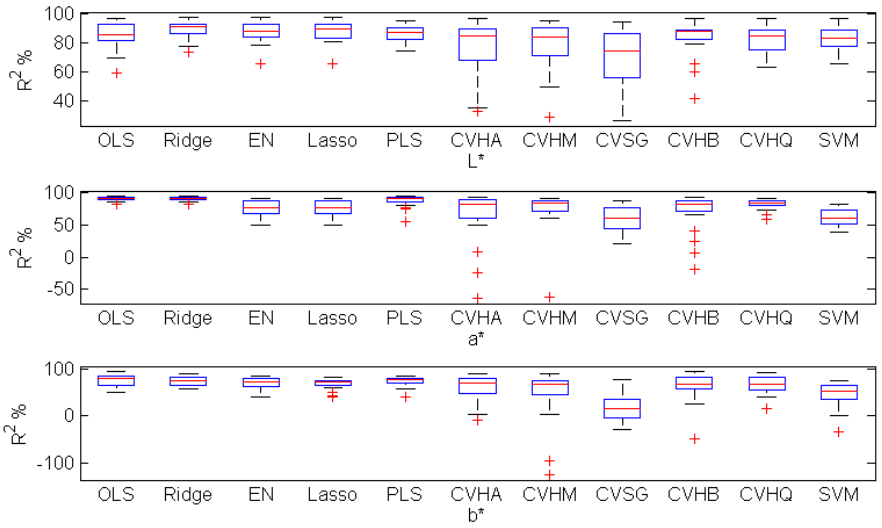


Figure A.13: R^2 box plots of the test data using EN-based Feature selection for linear, non-linear and SVM methods

proposed features. In the case of SVM results, the most prominent improvement obtained for the a^* component. Comparing all the results, the best ΔE_{ts} was 2.87 obtained from the ridge regression using the EN-based feature selection.

A.6.7 Comparison with RGB Images

In order to investigate the effect of the number of wavelengths in the accuracy of the regression models, we have extracted the RGB components from the 20 original bands. Then, these pseudo RGB features were used to perform $L^*a^*b^*$ prediction using the best linear and non-linear models from the previous experiments as well as the SVM method. The average results over the 25 data sets are presented in Table A.4. The prediction result in the case of L^* component, is good, showing that for brightness component, the use of three RGB bands may be enough. The results for the chromatic components are worse than the multispectral bands specially in the case of the b^* component. We can see that the complex non-linear methods can do significantly better predictions on the features from the limited RGB bands for the chromatic components, compared to the linear and kernel-based models. All the ΔE_{ts} values are above 4.

Table A.4: Average R^2 and ΔE_{ts} of the test data on pseudo RGB features

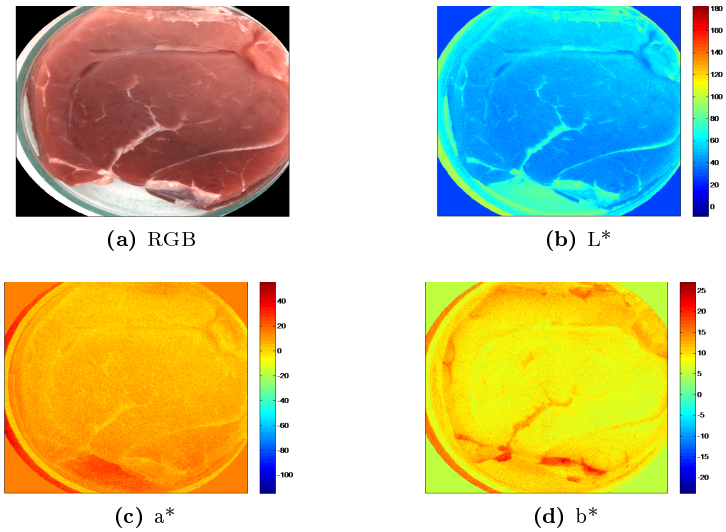
R^2	Ridge	EN	Lasso	CVHB	CVHQ	SVM
L*	87.53	87.34	87.60	88.13	86.48	87.10
a*	47.35	35.03	33.96	49.06	62.00	31.87
b*	14.02	9.11	10.33	28.87	20.84	7.95
ΔE_{ts}	4.68	4.95	4.92	4.38	4.30	4.99

Although a real RGB image captured by a CCD camera may not be exactly the same as the images we formed by band extraction over the multispectral images, the poor prediction results for the color components compared to the results using multispectral bands, can demonstrate the superiority of the multispectral imaging.

A.6.8 Displaying $L^*a^*b^*$ Components

In order to visualize the results of the $L^*a^*b^*$ color predictions, we made a prediction for all the pixels of a meat sample. To form these images, one of the trained ridge models on the EN-based feature selection method was used for each of the three components. Figure A.14 illustrates the pseudo RGB image and the corresponding images of the $L^*a^*b^*$ components. In the L^* image, the main structure of the marbled meat is distinguishable. In the a^* and b^* image, we can observe the color variation in different parts of the meat.

We investigated the use of multispectral images of raw meat for $L^*a^*b^*$ color prediction. Considering the variation in the results of the same methods on the 25 random sets, the important role of an appropriate training set, covering the existing variation of the population, in success of the prediction model becomes clear. Another point is that, comparison of the best results of different models show that, the use of a sub-set of features can improve the results. In our work, the proposed supervised linear feature selection algorithm outperformed the PCA for all tested methods. However, the best results were obtained by applying a non-sparse linear regression method like ridge on these features. SVM was the next best method for the selected features. Although the non-linear methods are more complex and more time-consuming in training, they did not obtain higher results in average, compared to the two other methods. Their box plots show that, an inappropriate initial point or a convergence to a local minimum may affect the final model dramatically and their average results may not improve due to these few poor outliers. On the other hand, the results show that more complex models work better on limited number of features. The $L^*a^*b^*$ predictions from pseudo RGB features support this.



(a) RGB **(b)** L*

(c) a* **(d)** b*

Figure A.14: The RGB image of a meat sample and its corresponding predicted $L^*a^*b^*$ components

In addition, we found that for prediction of the L^* component, simple RGB bands give good average result. But, they fails to gain acceptable results for the chromatic components.

Another important point in terms of the reduction in wavelengths is that, for each of the three components, the reduced number of wavelengths by the proposed method can perform an acceptable prediction. The best average test results of the all three strategies and their combination with the pre-processing methods are compared in Figure A.15. In addition, the comparison of the best ΔE_{ts} of these four approaches are presented in Figure A.16.

The selected features in Figure A.11, showed high frequencies in selection of the NIR wavelengths together with some visible bands in all cases. This shows the importance of the spectral imaging. In (Cao and Jun, 2008), the GRNN (CVSG) was suggested for CMYK color conversion into the $L^*a^*b^*$. Considering the tested non-linear models, we can see that in the case of multispectral images of meat, this model shows the lowest performance compared to the other models. However, there was no comparison in (Cao and Jun, 2008) between different ANN models. In (Larrain et al., 2008), the regression of colorimeter measurements on RGB images of only beef samples gained the highest R^2 for a^* component (96%), while for the two other components it was less than 60%. In our work, the best R^2 was also obtained for a^* component and the R^2 of the

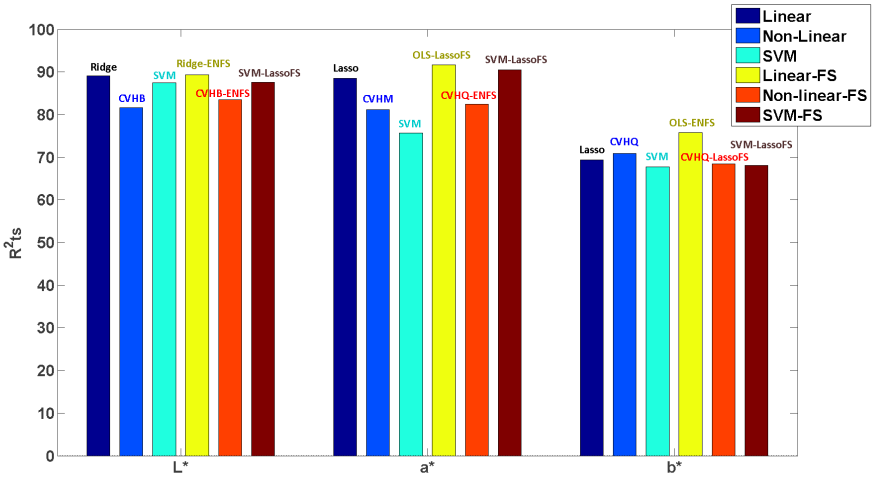


Figure A.15: Comparison of the best average R^2 test results for the linear, non-linear and SVM methods and their combinations with the feature selection (FS) methods.

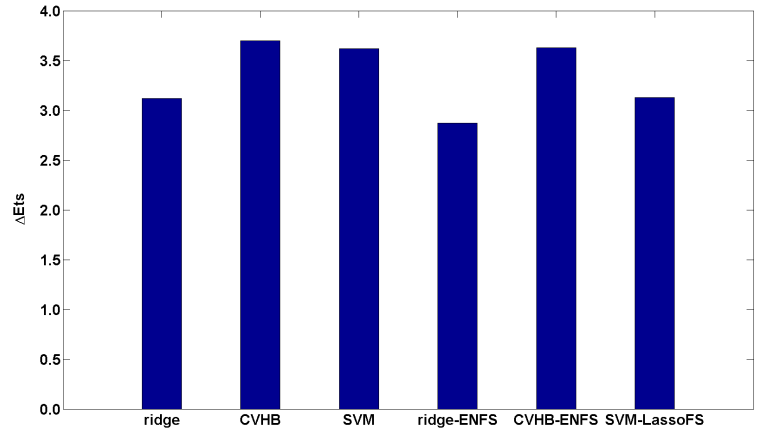


Figure A.16: Comparison of the best average ΔE_{ts} for the linear, non-linear and kernel-based methods and their combinations with the feature selection (FS) methods.

two other components was higher. However, the best ΔE_{ts} in our work was less than that work (2.87 and 1.57 respectively). The main reasons are the random division and averaging over 25 test sets and also the use of different meat types (veal, beef, chicken, pork, etc.) than one item, makes the fitting task with the prediction models more difficult.

We believe that, the mismatch between the regions where measurements were performed and the ROI regions are likely one main source of error in our models. In addition, as stated before, the random division of the original data set,with limited samples of many varieties, into training and test sets can be another source of error. Because it raises the possibility that some of the training sets do not cover the existing variability inside the original data set and therefore, the average results be decreased.

A.7 Conclusion

In this paper, multispectral images of different kinds of raw meat were used for prediction of the $L^*a^*b^*$ color components, which is useful for food quality inspection. The use of meat images was preferred over the use of standard color checkers due to the special characteristics of raw meat such as transparency and fiber structure. Results from the experiments supports this. Three regression strategies, linear, non-linear and kernel-based (SVM) were compared for color conversion. In addition, finding a sparse solution with a minimum number of wavelengths is important, since they are economically more effective for industrial vision systems. Therefore, a supervised linear feature selection algorithm was proposed. This method was compared with PCA using all three strategies. In order to generalize the results and make a reliable comparison between different methods, the original data set was randomly divided 25 times into training and test sets. Comparison of the results showed that the proposed feature selection strategy with non-sparse linear regression gained the best average results for all the color components. Finally, comparison with the pseudo RGB data showed the superiority of the multispectral data for prediction of the chromatic components.

Acknowledgment: This work was (in part) financed by the Center for Imaging Food Quality project which is funded by the Danish Council for Strategic Research (contract no 09-067039) within the Program Commission on Health, Food and Welfare.

Appendix B

DCT -Based Characterization of Milk Products Using Diffuse Reflectance Images

---

Authors: Sara Sharifzadeh[1], Jacob L. Skytte[1], Line H. Clemmensen[1], Bjarne K. Ersbøll[1].

1. Department of Applied Mathematics and Computer Science, Technical University of Denmark.

Published in proceedings 18^{th} *International Conference on Digital Signal Processing (DSP 2013), 1-3 July 2013, pp.1-6.*

Abstract

We propose to use the two-dimensional Discrete Cosine Transform (DCT) for decomposition of diffuse reflectance images of laser illumination on milk products in different wavelengths. Based on the prior knowledge about the characteristics of the images, the initial feature vectors are formed at each wavelength. The low order DCT coefficients are used to quantify the optical properties. In addition, the entropy information of the higher order DCT coefficients is used to include the illumination interference effects near the incident point. The discrimination powers of the features are computed and used to do wavelength and feature selection. Using the selected features of just one band, we could characterize and discriminate eight different milk products. Comparing this result with the current characterization method based on a fitted log-log linear model, shows that the proposed method can discriminate milk from yogurt products better.

Keywords:

discrete cosine transform; entropy; diffuse reflectance image; discrimination power.

B.1 Introduction

The Discrete Cosine Transform (DCT) is an appropriate transformation in the field of signal processing. It was first introduced in (Ahmed et al., 1974) to be used in the image processing area for the purpose of feature selection. It has excellent decorrelation properties as well as energy compaction. In addition, it decomposes the spatial frequency of an image in terms of various cosines transforms. Some of its application areas are image and speech compression (Gonzalez and Woods, 2001; Ramirez and Minami, 2003), speech recognition (Bouvrie et al., 2008; Sharifzadeh et al., 2012a) and medical imaging (Fu et al., 2005).

In this paper, the DCT is employed for decomposition of diffuse reflectance images. These images are obtained by illumination of a hyperspectral coherent laser (460-1000 nm) into the surface of eight different milk products. This vision system has been introduced recently for inspection of the structure of food items (Nielsen et al., 2011a,b). It is applicable for homogenous products where particle size and shape are important parameters. The main idea is to

use the diffusion effects, which are known to be correlated to the microstructure, for characterization of the structural composition of food items (Martelli et al., 2010; Mateo et al., 2010).

On the other hand, research findings in the field of food quality control have demonstrated a correlation between the texture, chemical and physical properties of food items with their microstructure characteristics (Aguilera, 2005; Bourne, 2002).

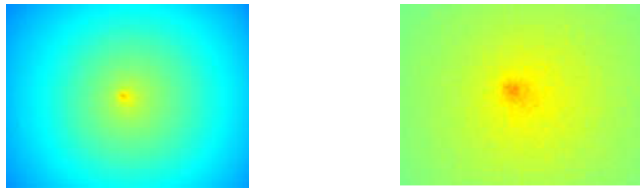
Considering these sequential relationships from the optical level to the quality level, it is possible to build an automatic light-based system as a measuring tool, for monitoring the quality of dairies along the production line and avoid unwanted structures during the process.

Therefore, finding an efficient method for characterization of the hyperspectral images into key discriminative features obtained from a minimum number of bands is of special concern in this field. The reduction in the number of required wavelengths will assist to simplify the laser set-up and make the overall system simpler and cost effective.

According to the characteristics of the milk products e.g. fat or viscosity, we can observe different visual effects in the hyperspectral images. The main optical feature is the low frequency light diffusion emanating from the incident point that has the highest intensity in the image as can be seen in figure B.1(a). Another important effect is a high frequency speckle pattern caused by interference of coherent light due to surface irregularities (Goodman, 2007). It is shown in figure B.1(b) by zooming in around the center point. These effects vary in different products according to their molecular composition and thus reflectance and scattering properties of light.

The current characterization technique for these images uses a narrow band of pixels of the scattering profile including the scattering center (Nielsen et al., 2011a,b; Sharifzadeh et al., 2012b). A double logarithm transformation is applied on the original profile to form this image. Therefore, the extracted line of intensities is called the log-log model. The resulting profile includes a slope and an intercept containing the subsurface and surface information respectively. This method only considers the low frequency information in the image.

In this paper, we propose to apply a DCT transform on the double logarithm of the entire diffuse reflectance image to decompose the low frequency diffusion effect as well as the high frequency speckle patterns. DCT can decorrelate the highly correlated information in these images. It decomposes the low frequency diffusion effects and high frequency speckle effects into low and high order coefficients that could be quantified easier. Finally, due to the high compression level



(a) (b)

Figure B.1: (a) A log-log transformed diffuse reflectance image of yogurt showing the low frequency diffusion effect at the center. (b) The zoomed image showing the high frequency speckle noise around the incident point caused by the destructive interference of light to the rough surface of fermented milk.

in the DCT domain, the number of discriminative features is reduced. In order to form an initial set of features for each image of each wavelength, we combine those of both low and high frequency effects. The low order DCT coefficients are considered to characterize the optical properties. The entropy information of the high order coefficients are used to characterize the speckle effect based on an approach that will be explained in section 3.

In the next step, the discrimination power analysis (DPA) introduced in (Dabbaghchian et al., 2010), is employed as a selection criterion on the initial set of features for both wavelength and feature selection. It is a more careful method in terms of discrimination than the conventional zigzag or zonal masking for DCT coefficient selection. Especially, that is in our work, both the low and high order features are important. Using the final selected features of one proper wavelength, we could characterize and discriminate the eight different products.

The proposed method is compared to the previous profile based characterization method including low frequency information and the results show that in addition to the more discrimination power of the proposed method (including both the low and high frequency information), it can separate the milk class products from the yogurt class better.

The rest of this paper is organized as follows. In section B.2, the data is described. Section B.3 presents the characterization of the diffuse reflectance images. In sectionB.4, feature selection and discrimination is explained. The experimental results are shown in section B.5. Finally, there is a conclusion for this paper.

Table B.1: The eight milk products and their fat levels

Product Type	Yogurt					Milk		
Short Names	L	M	H	CH	CU	L	M	H
Fat Level	0.5	1.5	3.5	0.1	1.5	0.5	1.5	3.5

B.2 Data Description

The data set consists of spectral diffuse reflectance images (1200×1600 pixel) of eight commercial dairy products including milk and yogurt categories. Table B.1 shows their names and fat levels. L, M and H stand for low, medium and high. The CH and CU are extracted from the commercial name of the products. In each category, there are products with different fat levels and viscosities. In the yogurt category, there are two different products with similar fat levels. The yogurt products differ from each other not only in terms of the fat, but also according to the applied fermentation processes. In this paper, we are not interested in predicting these kinds of features. Instead, we would like to characterize the products diffuse reflectance profiles and then discriminate them using their optical features. In fact, the optical characteristics represent the chemical, physical and structural differences between the products. For each product, there are five samples in the data set. Thus, there are 40 samples available in total. The laser was illuminated in 55 wavelengths (460-1000 nm).

B.3 Characterization of The Images

As mentioned in section B.1 there are two important features in the diffuse reflectance images that can be used for characterization of these images; the low frequency light diffusion effect and the high frequency speckle effect.

The light diffusion effect shows the spatial intensity distribution due to the absorption and scattering of the light. It is mostly dependent on the microstructural characteristics of the subsurface such as particle size distribution.

The speckle effect is caused by the interference of light at the surface. It can be seen as a measure of surface roughness. In a fermented milk product like yogurt, the surface roughness is higher than milk due to the increase in viscosity of the material after the fermentation process. Hence, it could be used as a measure for distinguishing milk from yogurt. Figure B.2 shows in the top row, two diffusion images of a medium-fat milk sample (M-M) and a high-fat yogurt (Y-

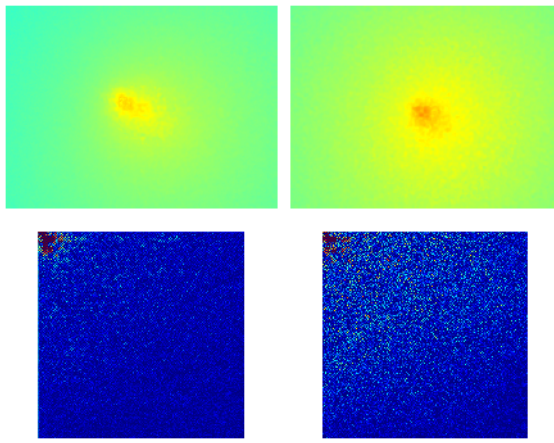


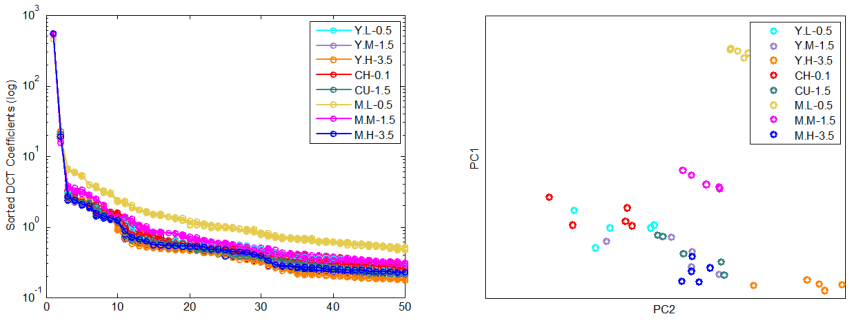
Figure B.2: (top) left and right, The zoomed diffused reflectance images of milk-M (1.5) and Yogurt-H (3.5) respectively. (Down) their corresponding 400 × 400 top left DCT coefficients from the DCT matrix.

H). The images are zoomed around the incident point. The difference in both low frequency diffusion effect and the high frequency speckle noise effect is clear between the two images.

B.3.1 DCT transform

Two dimensional DCT transform is applied to the diffuse reflectance images of each sample product at each wavelength. This yields 40×55 DCT matrices of size 1200 × 1600. In the second row of figure B.2, the corresponding 400 × 400 DCT coefficients from the top-left DCT matrix of the above images are illustrated. The difference in the higher order DCT coefficients represents the speckle effect that was seen in the spatial domain as well. However, it is not easy to distinguish the difference in low order DCT coefficients that represent the diffusion effect.

According to these observations, choosing the DCT coefficients in a conventional zigzag or zonal low order masking alone, is not a good choice. That is due to the large number of DCT coefficients in a wide span of low and high frequencies that describe the scattering and speckle effect. To demonstrate this issue, a 400×400 sub-volume of DCT coefficients from the top-left of the DCT matrix is considered for all the samples of all classes. For ease of visualization, they are



(a) (b)

Figure B.3: The first 50 highest DCT coefficients of all the samples of the 8 products: (a) in original domain (b) in PCA space using the first two PCs.

sorted and just the logarithm of the 50 highest are illustrated in figure B.3(a). It is difficult to distinguish all the products. In addition, they are transformed into the PCA domain and the first two PCs are shown in figure B.3(b). In both images, just a few products can be distinguished from each other and the other classes. It is not easy to distinguish most fermented products and the high-fat milk from each other. This is because; the higher values of the DCT coefficients only carry the information about the diffusion effect and that is not enough for discrimination. In order to include the speckle effect, we propose to use the entropy of the DCT coefficients which will be explained in the following section.

B.3.2 Entropy

The high frequency DCT coefficients that contain information about the speckle effect result in an increase in the entropy of the sub-volumes of the DCT matrix that include them. For example, in the two 400×400 sub-volumes that are shown at the bottom of figure B.2, the entropies are 1.55 and 2.02 from left to right respectively. Starting from the top-left corner of a DCT matrix, we considered an $n \times m$ sub-volume and calculated the entropy repeatedly, while continuously increasing the n and m values as illustrated in figure B.4(a). The resulting entropy profile is shown in figure B.4(b). It shows that, as the size of the volume increases, the entropy also increases up to some point and then, decreases due to the uniform values of the DCT coefficients in higher frequencies. Since the speckle effect that characterizes the surface roughness enhances the higher order DCT coefficients, the maximum entropy should describe the speckle

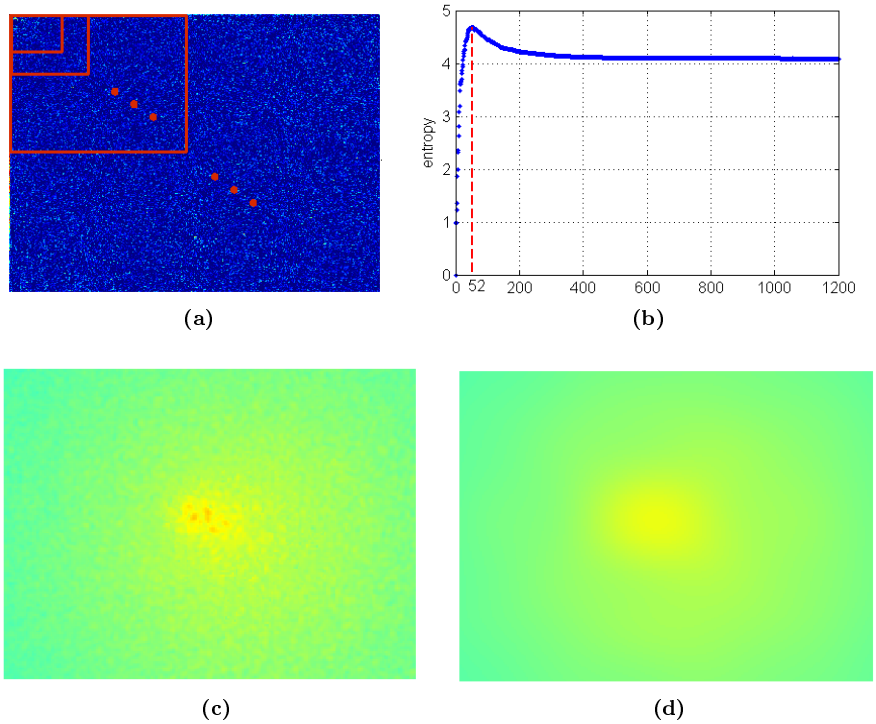


Figure B.4: (a) The sequential entropy calculation on increasing sub-volumes of the DCT matrix. (b) The resulting entropy profile. (c) The zoomed original diffusion image around the incident point. (d) The diffusion image obtained by the inverse DCT transform of the 52×52 lower order sub-volume of the DCT matrix .

effect for each sample. By forming such entropy profile for the eight products, we found that it can characterize their speckle effect uniquely. Therefore, the low entropies before the peak point can be considered as the diffusion effect so that, their corresponding sub-volumes include mostly the DCT coefficients describing the diffusion effect. On the other hand, the right side of the peak point includes the higher order DCT coefficients that describe the diffusion effect. To verify this further, we isolated the low order diffusion effect DCT coefficients using the index of the peak point that is 52 in figure B.4(b). Then, an inverse DCT transform is applied to this 52×52 DCT sub-volume. Comparison of the result with the original diffuse reflectance image shows the removal of the speckle effect, as shown in figure B.4(c, d).

B.3.3 Forming the Initial feature set

According to the discovered points, the right side of the entropy profile was considered for characterization of the speckle effect. The mean, the standard deviation and the maximum value, of this part of the profile were considered as the candidate speckle effect features. By looking to the entropy profiles of the eight products, it was found that in average, the maximum entropy occurs around a 50×50 sub-volume. Regarding to its variation in different products and also considering a softer threshold for separation of the DCT coefficients of the diffusion and speckle effects, a 20×20 sub-volume of low order DCT coefficients was considered. They form a 400 length vector as the candidate feature for the light diffusion effect.

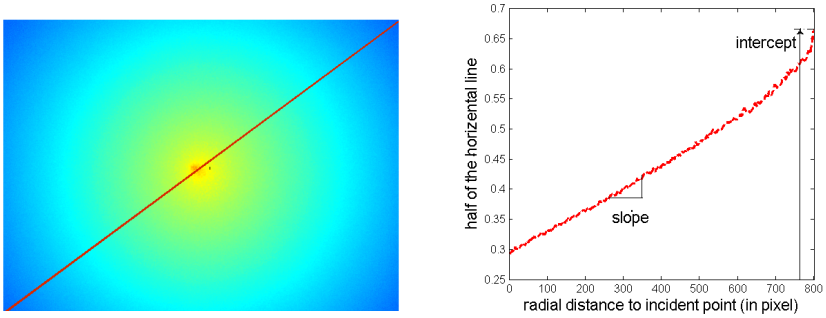
The final initial set of features for each wavelength image was formed by concatenating the three candidate features of the speckle effect with the 400 of the diffusion effect.

B.3.4 Feature forming based on log-log model

In order to form the features based on log-log model, at each wavelength, a narrow diagonal band (around 10 pixels width) including the scattering center was considered in the double logarithm of the diffuse reflectance image as shown in figure B.5(a). The orientation of the line was chosen in a way to consider as much as possible, higher number of pixels along the path through the center. Then, it was averaged over the pixels. Since this diagonal line is symmetric, just half of that was considered. The resulting averaged profile includes an intercept from the peak and a slope as shown in figure B.5(b). These two features were used to characterize the image. For more information, we refer to (Nielsen et al., 2011a,b).

B.4 Feature Selection and Discrimination

The length of the formed initial set of features (403) per band, regarding the total number of samples of all classes (40) is quite high. Therefore, it is better to select a subset of them according to their ability for characterization and discrimination of different products. Besides that, there are 55 bands per sample and as mentioned earlier, we are interested to reduce the number of wavelengths to simplify the laser set-up. Therefore a strategy should also be taken into



(a) (b)

Figure B.5: (a) Symmetric narrow band of pixels crossing the incident point in the double logarithm diffuse reflectance image. (b) Half of the band is averaged and the slope and intercept from the peak are shown.

account to sort the discrimination ability of different wavelengths and select one or a few number of them.

Since majority of the features are the decorrelated DCT coefficients, it is not necessary to decorrelate them by a transformation into an orthogonal space. Inspired by the approach in (Dabbaghchian et al., 2010), we employ the DPA introduced in that work. The main idea behind this data-dependent approach is that, all of the DCT coefficients do not have the same discrimination power (DP). In other words, some of them can discriminate the classes better than the others. It is different from other similar approaches such as PCA and LDA, in the sense that it does not utilize the between- and within- class variances by a transformation to maximizes the discrimination of the features in the transformed domain. It searches for the best discriminant features in the original domain. In case of decorrelated features such as DCT coefficients it is an appropriate approach for ranking the features and choosing a sub-set of them. The calculation of DPA will be explained step by step in the following:

Assuming that we have C classes with the N_c number of data points and $P = 403$ features in each class, the DP_j of each feature f_j $(j = 1, 2, ..., 403)$ is calculated as follows:

1. The mean and variance of each class is calculated for that feature (f_j) : $m_{jc} = \frac{1}{N_c}\sum_{n=1}^{N_c}(f_{nj}), C = 1, 2, ..., C$, $v_{jc} = \sum_{n=1}^{N_c}(f_{nj} - m_{jc})^2, c = 1, 2, ..., C$
2. The variance of all classes are averaged: $V_j^W = \frac{1}{C}\sum_{c=1}^{C} v_{jc}$

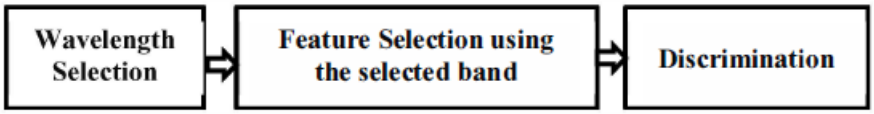


Figure B.6: The three steps of the sequential strategy

3. The mean and variance of all training samples are calculated for $f_j : M_j = \frac{1}{C \times N_c} \sum_{c=1}^{C} \sum_{n=1}^{N_c} f_{nj}$, $V_j^B = \sum_{c=1}^{C} \sum_{n=1}^{N_c} (f_{nj} - M_j)^2$
4. The DP can be estimate as $DP_j = \frac{V_j^B}{V_j^W}$

It is mentioned in (Dabbaghchian et al., 2010) that DPA can be used as a stand-alone feature reduction algorithm. Since we need to do both band and feature selection, a sequential strategy is taken into account as shown in figure B.6.

B.4.1 Preparation of training and test sets

In order to maintain the training and test sets from the few data points, one sample of each class was considered as the unseen test data and the rest were assigned to the training set. Therefore, the two sets were formed as $tests_{8x403X55}$, $train_{32X403X55}$. Then, leave one out cross validation (LOOCV) was used on the training data set for both wavelength selection and feature selection steps. LOOCV is used for generalization and to avoid over-fitting as much as possible (Hastie et al., 2009). However, due to the limited training data points, this could not be achieved completely.

B.4.2 Wavelength Selection

The band selection algorithm is as follows:

1. At each iterations of LOOCV, sum of the DPA of all 403 training features are calculated at each wavelength ;$w = 1, 2, ..., 55$; $SUM_{32\times55}$.
2. The sum of DPs, $SUM_{32\times55}$is averaged over the 32 iterations; $Average - SUM_{1\times55}$.
3. The best band is the one with the highest average discrimination power. This algorithm was also used for wavelength selection of the log-log model.

B.4.3 Feature selection for the selected band

The use of just one band is a significant reduction in the number of features, since there are 403 initial features per wavelength. In order to select the most discriminative features of the selected band, these steps are followed:

1. The DPs o f the features in the selected band are sorted for each of the 32 Looev iterations in descending order. Then, the corresponding features to the first top five DPs at each iteration are kept in a list; $list_{32\times5}$

2. The densities of the N_u unique features in this list are calculated. $Density_{1\times Nu}$

3. According to these densities, the features that were among the top five features almost in all 32 LOOCV iterations are selected as the final features. The number five in the above explained procedure was chosen empirically by looking to the sorted features and also for the aim of selecting a limit number of features. Interestingly, we observed in all the iterations, the first three features were from distinct low frequency DCT coefficients representing the light diffusion effect and one of the last two was the mean value of the speckle effect from the entropy profile shown in figure B.4(b).

B.4.4 Discrimination

In order to evaluate the proposed characterization approach and compare it with the existing log-log method, the training and test data are visualized on the same plot. Besides that, the discrimination power of the two methods is numerically measured by sum of the feature's DPs as well as the maximum Rayleigh quotient term (Hastie et al., 2009):

$$\max \frac{a^T B a}{a^T W a} \tag{B.1}$$

Where B and W are the between- and within-class covariance matrices and a is the Eigen vector of the generalized Eigen value problem, $det(B - \lambda W) = 0$. In order to maximize equation B.1, the Eigen vector a_i corresponding to the highest Eigen value λ_1 should be used. In addition, the support vector machine (SVM) classifier with a linear kernel is used (Chang and Lin, 2011) and the average LOOCV results and unseen test results are compared for the two methods.

B.5 Results and Discussion

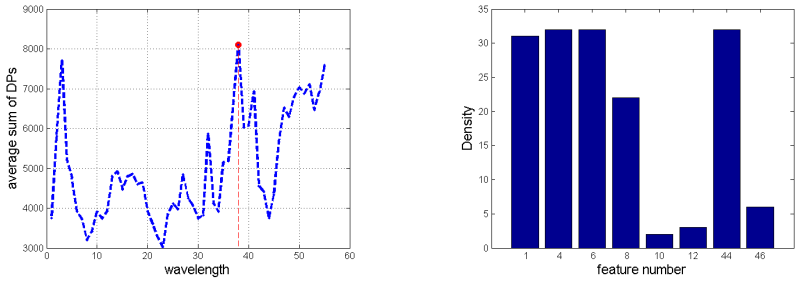
First, the results of the proposed method in DCT domain will be shown. Then, the log-log model results will be presented. Finally, there is a discussion.

B.5.1 Characterization results in DCT domain

As explained in the previous section, both the band selection and feature selection were performed using LOOCV on the training data. Figure B.7(a) shows the average sum of the DPs for the 55 bands. According to this plot, the highest sum of DPs obtained for band 38 (830 nm).

By sorting the feature's DPs in this band, a list of features corresponding to the top five DPs were formed for the 32 LOOCV iterations. There were eight unique features in the list. Figure B.7(b) shows the densities of the unique features in the list. As can be seen, three features were among the top features in all 32 iterations. They are the low order DCT coefficients showing the light diffusion effect. Their location in the DCT matrix is represented in figure B.8. In addition, the feature number one that represents the mean entropy of the speckle effect was among the top five features in 31 of the iterations. These four features were selected as the final features, for characterizing the samples.

In order to visualize the ability of the speckle effect features to separate the two groups of yogurt products and milk, a 3D visualization of the mean, standard deviation and maximum features (1, 2, 3 in figure B.8) is represented in figure B.9(a). The results show that these features are capable to perform the separation accurately for both training and test data. In addition to this between group separations, we can also observe a trend for within group separation according to the fat level. Figure B.9(b) shows the 3D visualization of the three diffusion effect features (4, 6, 44 in figure B.9). As can be seen, they fail to separate the high-fat milk sample (M-H) and the medium-fat yogurt (YM). Since the visualization of the 4D selected features is impossible, three of them (1, 4, 6) are chosen and visualized in figure B.9(c). Even in absence of one of them, we can see the successful separation of all the classes and also the two groups of milk and yogurt. Finally, the four features are transformed into the orthogonal PCA space and the first two PCs are shown in figure B.9(d). Besides the successful discrimination, we can observe that the $PC1$ represents the variation from yogurt to milk group, while $PC2$ shows the change in fat content.



(a) (b)

Figure B.7: (a) Average sum of DPs over the 32 LOOCV iterations for the 55 bands. (b) Density of the 8 unique features found among the top five discriminative features in the list over the 32 iterations. The horizontal axis shows the feature's number among the 403 features.

B.5.2 Characterization results using the log-log model

The same wavelength selection strategy based on sum of discrimination powers were used for band selection for log-log model dataset. Figure B.10 shows the 2D visualization of the slope and intercepts features in original as well as PCA space. In both spaces, the two features group the samples only according to their fat level, while there is no trend to separate the milk group from the yogurt group. For example the high fat milk (M-H) and the medium fat yogurt (Y-M) have close overlap which may make the discrimination difficult.

B.5.3 Discussion

According to the visualized results, the combination of the speckle effect (high frequency) and diffusion effect (low frequency) features in DCT domain shows to be a promising way of characterizing the diffuse reflectance images. The statistical analysis results are presented in table B.2. Although both methods could discriminate the single test samples of all classes, the average LOOCV classification performance shows that the proposed method can work better. However, the statistical models suffer from the over-fitting due to the limited number of samples. The table results show that, the DCT domain features are capable to characterize the images better in terms of discrimination power and Rayleigh criteria than the log-log model features. Besides that, considering the plots in figure B.9 and figure B.10, they are capable to reduce the overlap between

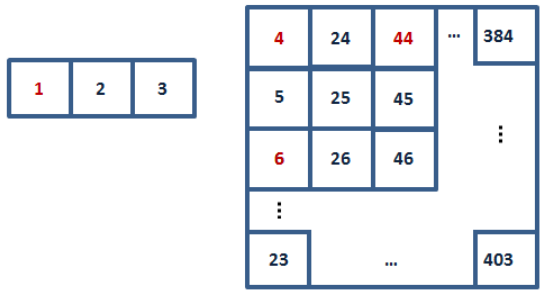


Figure B.8: The 1, 2 and 3 are the mean, standard deviation and maximul12 of the entropy profile of the speckle effect. The 400 low order DCT feature's numbers start from 4.

Table B.2: The discrimination results

	Av. SVM Prf. of LOOCV	SVM Test Perf.	Sum of DPs	Rayleigh criteria
DCT	100%	100%	2460.3	5850.6
Log-Log Model	96.87%	100%	1815.1	79.60

classes and separate the products not only according to their fat level, but also according to their category (milk-yogurt). That is obtained by employing the ability of DCT transform in frequency decomposition and combining the high and low frequency information of the images. When only the analysis of the diffusion effect is needed, this frequency decomposed information can be used to exclude the speckle effect as shown in figure B.4(d), using the inverse DCT transform.

B.6 Conclusion

In this paper, a DCT-based characterization method is introduced for diffuse reflectance images. These images result from illumination of a narrow laser beam in different wavelengths into eight different dairies. They were milks and yogurts of different types and fat levels. The low order DCT coefficient were used to characterize the low frequency light diffusion effect and the entropy information of higher order DCT coefficients were used to characterize the speckle effect in the images. The discrimination power criterion was used to reduce the number of wavelength and to select the features. The existing characterization method based on a linear log-log model can only separate the products according to

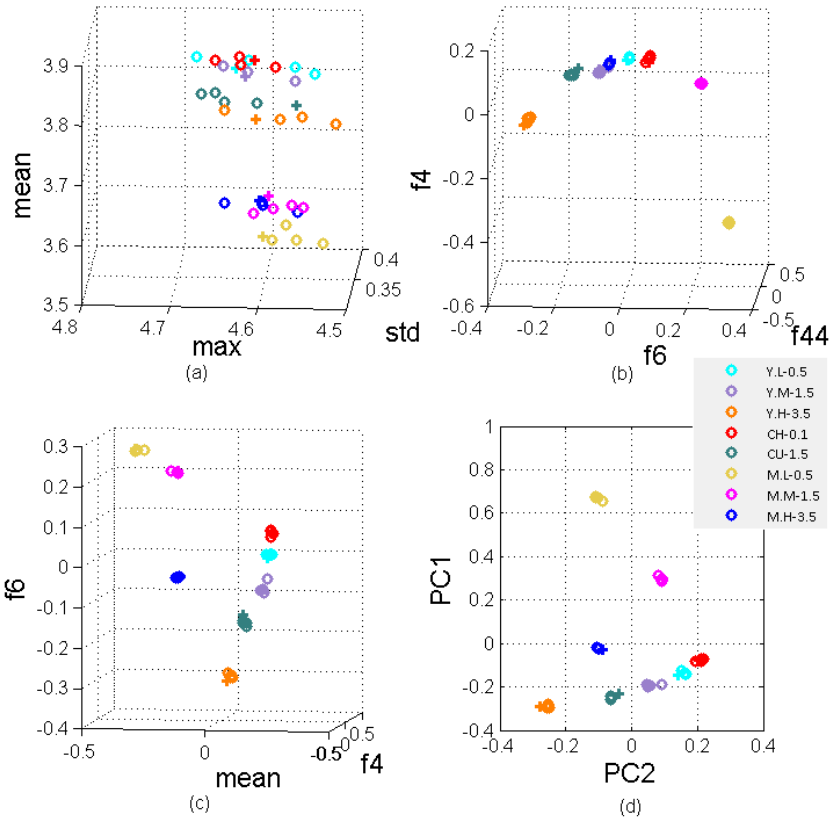


Figure B.9: (a) 3D visualization of speckle effect features (b) 3D visualization of the three diffusion effect selected features (c) 3D visualization of speckle and diffusion selected features (d) 2D plot in PCA space using the first two PCs. The "O" shows a training sample and "+" shows a test sample.

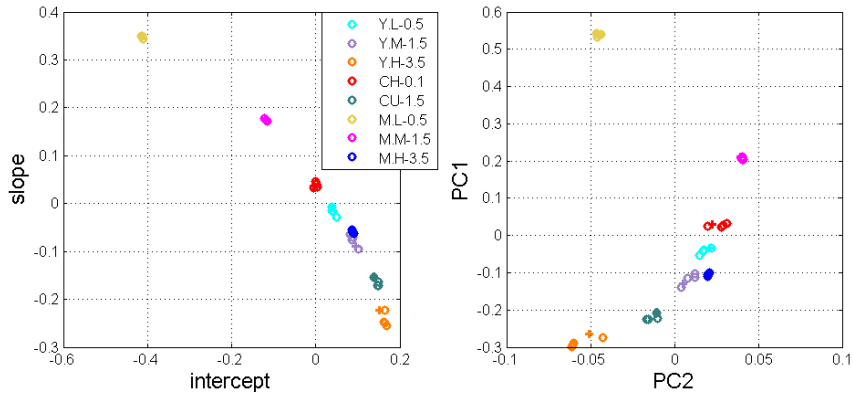


Figure B.10: 2D visualization of the log-log model features (a) in original space (b) in PCA space

their fat levels, but the proposed method can discriminate them based on both their category (milk-yogurt) and fat level. It also improves the discrimination and removes the overlap between the classes.

Acknowledgment: This work was (in part) financed by the Center for Imaging Food Quality project which is funded by the Danish Council for Strategic Research (contract no 09-067039) within the Program Commission on Health, Food and Welfare.

Appendix C

Sparse supervised principal component analysis (SSPCA) for dimension reduction and variable selection

---

Authors: Sara Sharifzadeh[1], Ali Ghodsi[2], Line H. Clemmensen[1], Bjarne K. Ersbøll[1]

1. Department of Applied Mathematics and Computer Science, Technical University of Denmark.
2. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON., Canada.

Submitted.

Abstract

Principal component analysis (PCA) is one of the main un-supervised pre-processing methods for dimension reduction. When training labels are available, supervised PCA is a better solution. In cases where both dimension reduction and variable selection are required, sparse PCA (SPCA) methods are preferred. In this paper a pre-processing method for sparse supervised PCA (SSPCA) is proposed. The method is based on the objective function of a supervised PCA algorithm and a punishment term is added to make the Eigen vectors sparse. To solve the new objective function, the penalized matrix decomposition (PMD) algorithm is employed. The PMD algorithm was used for a SPCA method previously. However, the proposed method achieves a higher level of sparsity compared to the PMD-based SPCA. SSPCA can be used for data sets with linear as well as non-linear behavior. The proposed method is compared with PCA, PMD-based SPCA and supervised PCA. Since there is similarity in the objective function of SSPCA and sparse partial least squares (SPLS) method, the results are also compared with SPLS. Experimental results from the simulated as well as real data sets show that SSPCA provides an appropriate trade off between accuracy and sparsity. Comparison of the results with the other methods show that in terms of accuracy it is one of the most successful methods, while in terms of sparsity, it performs excellent variable reduction. Therefore, the Eigen vectors found by SSPCA can be used for feature selection in different applications.

Keywords:Variable selection, Dimension reduction, sparse PCA, supervised PCA, sparse supervised PCA, penalized matrix decomposition

C.1 Introduction

Principal component analysis (PCA) is a well known dimension reduction approach that is used in many data mining and machine learning problems such as genetics, image and signal processing, chemistry, etc. Given a data matrix $X_{N\times P}$ with N data points and P features, it maps data into an orthogonal space based on the sorted variance of the input data. In the new space, each principal component (PC) is a linear combination of all original variables. The first principal component corresponds to the highest variance and the second to the second highest variance and so on.

However, based on the type of problem, two main limitations can be considered for PCA; First, is that PCA is not sparse, while in many applications, specially

those with a high number of variables, it is important to reduce the number of variables and remove any irrelevant or noisy variable. For example, in spectral imaging applications, each variable might be a wavelength and sparse PCs result in a simpler vision set-up or in biology, each variable might correspond to a specific gene and interpretation of the sparse PCs are easier. This also makes it possible to employ any suitable non-sparse data analysis method afterward. In addition, PCA is un-supervised. Although this can be considered as an advantage in many cases, it can also be a limitation when a label or response vector is available because, it is not possible to guide the algorithm based on the target response. This is specially important when the task is regression or classification, where it is preferred to map data into a space based on the data variations that depend on the response and not necessarily according to the maximum variation.

To address the first limitation, many researchers have proposed methods and algorithms for sparse PCA (SPCA). Simple thresholding of the loadings was proposed in (Cadima and Jolliffe, 1995). In (Sigg and Buhmann, 2008) the sparse and non-negative PCA problem was addressed based on constraints on cardinality and sign of the elements. In (B. Moghaddam, 2006; A. d'Aspremont and Ghaoui, 2007), greedy algorithms were used to find sub-optimal solutions for SPCA. Another algorithm called SCoTLASS based on regression or reconstruction error property of PCs was developed in (Jolliffe et al., 2003). In (Zou et al., 2004), an SPCA algorithm was proposed using the Elastic-Net framework for L_1-norm penalized regression on regular PCs using least angle regression (LARS). In (Witten et al., 2009) an algorithm based on the penalized matrix decomposition (PMD) was proposed. The augmented Lagrangian method (ALSPCA) by (Lu and Zhang, 2009), regularized singular value decomposition (SVD) method by (Shen and Huang, 2008) and the generalized power method by (Journée et al., 2010) are alternative methods for computing the sparse PCs. Most of the solutions to SPCA are non-convex optimization procedures that find a solution close to the optimal point. Some of them such as (d'Aspremont et al., 2007; Zhang and Ghaoui, 2011) also guarantee the global convergence. In (d'Aspremont et al., 2007) an algorithm called DSPCA based on semidefinite programing (SD) was proposed by semidefinite relaxation of the SPCA problem. The second work (Zhang and Ghaoui, 2011), is based on the DSPCA for large scale data sets. First, a feature elimination method was used to reduce the problem size and then, a block coordinate descent algorithm, was used to solve the DSPCA. In (Vu et al., 2013) a convex relaxation of sparse principal subspace was proposed based on the convex hull of rank-d projection matrices (Fantop). The solution was based on SD and generalizes the DSPCA approach to d $\geq$ 1 dimensions. Recently, a two-stage sparse PCA procedure has been proposed that attains the optimal principal subspace estimator in polynomial time (Zhaoran Wang, 2014). In another work, a robust algorithm for SPCA was proposed which is resistant to outlying observations (Croux et al., 2013).

For supervised dimensionality reduction, various approaches exists such as metric learning and sufficient dimension reduction methods (Chang and Yeung, 2006; Yeung and Chang, 2006; Torkkola, 2003; Fukumizu et al., 2004). In addition, supervised PCA methods were proposed (Bair et al., 2006; Barshan et al., 2011). In (Bair et al., 2006), a pre-processing step was added to conventional PCA. So that, based on the regression coefficients of initial features, only a subset of features with higher scores are considered for PCA. The supervised PCA method proposed in (Barshan et al., 2011) is a generalization of PCA which aims at finding the PCs with maximum dependency to the response variables. In that work, the Hilbert–Schmidt independence criterion (HSIC) (Gretton et al., 2005) was used as the dependency function between the data and target response. A closed form solution was found for the objective function.

This work is focused on developing a sparse supervised PCA (SSPCA) algorithm. Such an algorithm will be appropriate for pre-processing of data sets for which a target response is available and a sparse solution for variable selection or interpretation is desired. The supervised PCA algorithm from (Barshan et al., 2011) is used to form an initial objective function. In order to find sparse solutions, penalization constraints for the Eigen vectors are considered. The resulting optimization problem is bi-convex and solved using the PMD algorithm (Witten et al., 2009). Due to the use of a kernel in the objective function, the solution can handle data sets with linear as well as non-linear behavior. The sparse Eigen vectors found by the SSPCA algorithm can be used either for projection of a data set or feature selection. The projection is based on maximum dependency of the data to the target instead of its maximum variation. In this paper, SSPCA is compared with PCA, the SPCA based on PMD algorithm and the supervised PCA method. The SSPCA objective function is close to the objective function of sparse partial least squares (SPLS) algorithm. Therefore, a comparison is also performed with SPLS. The experiments were conducted on both simulated and real data sets.

The rest of this paper is organized as follows; In section C.2, the supervised PCA and SPCA based on the PMD method are explained. Section C.3 introduces the SSPCA method. The SPLS method is explained briefly in section C.4. Experimental results are presented in section C.5. Finally, discussion and conclusion are given in sections C.6 and C.7 respectively.

C.2 Related works

C.2.1 Supervised PCA

Considering a data matrix $X_{n\times p}$ that has n data points and p features, and also a target vector $Y_{n\times 1}$, supervised PCA finds a sub-space XV such that the dependency between the projected data XV and the outcome Y is maximized (Barshan et al., 2011). The HSIC independence criterion (Gretton et al., 2005) was used to measure the dependency.

According to HSIC, the independence of the variables X and Y is possible, if and only if any bounded continuous function of them is uncorrelated. Therefore, dependency and correlation are different. If two random variables are independent, their HSIC value will be zero. The HSIC can be expressed in terms of kernel functions. In (Barshan et al., 2011), an empirical form of HSIC was used to make it a practical criterion for independence testing. Let $Z = (x_1, y_1), \ldots, (x_n, y_n) \subseteq X \times Y$ be a series of n independent observations drawn from $P_{X,Y}$. The empirical estimate of HSIC is:

$$HSIC(Z, F, G) = (n-1)^{-2} tr(KHLH), \tag{C.1}$$

where F and G are separable reproducing kernel Hilbert space (RKHS), containing all continuous bounded real-valued functions of x and y respectively (from X to $\mathbb{R}$ and from Y to $\mathbb{R}$), k and l are the corresponding kernels of F and G and $H, K, L \in \mathbb{R}^{n\times n}, K_{ij} = k(x_i, x_j), L_{ij} = l(y_i, y_j)$ and $H_{ij} = I - n^{-1}ee^T$ is the centering matrix (e is a vector of all ones). Therefore, in order to maximize the dependency between two kernels, the value of the empirical estimate, i.e., $tr(KHLH)$ should be maximized. Then the objective function of supervised PCA is:

$$\max_V tr(KHLH) = \max_V tr(HXVV^TX^THL) = \max_V tr(V^TX^THLHXV), \tag{C.2}$$

where V is the orthogonal transformation which maps data into a new space where the features are independent. Thus, for supervised PCA, the following optimization problem was solved in closed form using Eigen vector decomposi-

tion:

$$\arg\max_{V} tr(V^T X^T HLHXV) = \arg\max_{V} tr(V^T QV), \tag{C.3}$$

$$\text{s.t.} VV^T = I.$$

If $Q = X^T HLHX$ is a symmetric and real matrix, with Eigen values $\lambda_1 \leq ... \leq \lambda_p$ and the corresponding Eigen vectors $v_1, ..., v_p$, then the maximum value of this cost function is $\lambda_p + \lambda_{p-1} + ... + \lambda_{p-d+1}$ and the optimal solution is $V = [v_p, v_{p-1}, ..., v_{p-d+1}]$. d is the dimension of the output space S.

In cases where $n \ll p$, the Eigen vectors of the very large $mathrmCov_{p\times p}$ should be calculated. That is impractical and therefore a dual form was proposed in (Barshan et al., 2011):

$$L = \Delta^T \Delta \Rightarrow Q = X^T HLHX = \Psi^T \Psi, \tag{C.4}$$

where $\Psi_{n\times p} = \Delta^T HX$ and V can be calculated by the SVD of $\Psi = U\Sigma V^T$. Then any training or test data can be transferred into the new space as $Z = XV$.

C.2.2 Sparse PCA

In a sparse PCA problem, the Eigen vectors should have some zero elements. This can be achieved by a penalization approach such as an upper constraint on the Eigen vectors as proposed in (Witten et al., 2009). If $X_{n\times p}$ be a data matrix of rank $K \leq min(n, p)$, an SVD problem is shown in Eq. C.5.

$$\hat{X} = U\Lambda V^T, U^T U = I_n, VV^T = I_p, \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_K > 0 \tag{C.5}$$

For $r \leq K$, the SVD problem was considered based on the Frobenius norm in (Witten et al., 2009):

$$\sum_{k=1}^{r} \lambda_k u_k v_k^T = \arg\min_{\hat{X}\in M(r)} \left\| X - \hat{X} \right\|_F^2 = \arg\min_{\hat{X}\in M(r)} \left\| X - U\Lambda V^T \right\|_F^2, \tag{C.6}$$

where $M(r)$ is the set of rank-r $(n \times p)$ matrices. In the case of the Frobenius norm, the following was demonstrated in the appendix of (Witten et al., 2009):

$$\frac{1}{2}\left\|X - U\Lambda V^T\right\|_F^2 = \frac{1}{2}\|X\|_F^2 - \sum_{k=1}^{K} u_k^T X v_k d_k + \frac{1}{2}\sum_{k=1}^{K} d_k^2. \tag{C.7}$$

Therefore, the minimization in Eq. C.6 was written as a maximization form for $k = 1$. In addition, by penalizing the decomposition of the original matrix X (PMD), sparse components U and V were achieved, as shown in Eq. C.8. The constant terms of Eq. C.7 were ignored.

$$\max_{u,v} u^T X v, \text{s.t.} \|u\|_2^2 \leq 1, \|v\|_2^2 \leq 1, P_1(u) \leq c_1, P_2(v) \leq c_2 \tag{C.8}$$

P_1 and P_2 are convex penalty functions like lasso $P_1(u) = \sum_{i=1}^{n} |u_i|$. The equality constraint on $\|.\|_2$ was changed into an inequality to avoid a non-convex problem. The objective function is bi-convex in u and v. That is, with u fixed, it is linear in v, and vice versa. Algorithm 3 shows the procedure for solving this bi-convex problem. That is the general $PMD(L_1, L_1)$ with penalty functions for both u and v.

Algorithm 3 Computation of K-factors of PMD

1. Let $X^1 \leftarrow X$
2. For $k \in 1, ..., K$:

(a) Find u_k, v_k and d_k by applying the following single-factor PMD algorithm to X^k :

- Initialize v_k to have L_2-norm equal to one.
- Iterate until convergence:
$$\begin{cases} u_k \leftarrow & \arg\max_{u_k} u_k^T X^k v_k, \\ & \text{s.t.} P_1(u_k) \leq c_1 \text{and } \|u_k\|_2^2 \leq 1 \\ v_k \leftarrow & \arg\max_{v_k} u_k^T X^k v_k, \\ & \text{s.t.} P_2(v_k) \leq c_2 \text{and } \|v_k\|_2^2 \leq 1 \end{cases}.$$
- $d_k \leftarrow u_k^T X^k v_k$.

(b) $X^{k+1} \leftarrow X^k - d_k u_k v_k^T$

The optimization equations in Algorithm 3 have a closed form solution based on soft thresholding. The parameters c_1 and c_2 are restricted to $1 \leq c_1 \leq \sqrt{n}$

and $1 \leq c_2 \leq \sqrt{p}$. The smaller the c_1 and c_2 values, the more sparse the Eigen vectors. For further demonstrations we refer to (Witten et al., 2009).

SPCA was solved as a $PMD(., L_1)$ problem in (Witten et al., 2009) which results in sparse column vectors. In addition, an orthogonality constraint was added to enforce orthogonality to the subsequent sparse PCs. There are similarities between this method and the work in (Shen and Huang, 2008) for identifying sparse principal components.

C.3 The proposed SSPCA method

In order to make the supervised PCA algorithm sparse, the Eigen vectors v_k are constrained in Eq. **??**. Then, the new objective function is:

$$\arg\max_{V}(tr(V^T Q V)) = \arg\max_{V}(tr(V^T \Psi^T \Psi V)) \tag{C.9}$$

$$\text{s.t. } \|v_k\|_1 \leq c_2, \|v_k\|_2^2 \leq 1.$$

The penalization constraint is applied on individual Eigen vectors and the same value is used for all Eigen vectors. Because Penalizing the Eigen vector's matrix and finding all the Eigen vectors simultaneously requires different regularization parameters for Eigen vectors, otherwise the resulting sparse matrix will be of rank one. That means an increase in the number of parameters which makes the problem more difficult. Therefore, in our work, we consider the same regularization parameter for all Eigen vectors and solve the problem for each Eigen vector separately. Then, there exist mathematical solutions for this simplified problem.

The equivalent SVD problem to this objective function is considered so that, $\Psi = U\Sigma V^T$. This transfers the problem into the form of PMD that finds the Eigen vectors in individual iterations. Since the column vectors, v_k must be sparse, it can be solved as an $PMD(., L_1)$ problem. That is, the penalization is applied only on column vectors.

$$\max_{u_k, v_k} u_k^T \Psi^k v_k \text{s.t. } \|v_k\|_1 \leq c_2, \|u_1\|_2^2 \leq 1, \|v_k\|_2^2 \leq 1, \tag{C.10}$$

$$u_k \perp u_1, ..., u_{k-1}.$$

Algorithm 4 Procedures for SSPCA

Input: training data matrix **X**, test data **x**, kernel matrix of target variable **L** and training data size **n**.

Output: Dimension reduced training and test data using sparse Eigen vectors, **Z** and **z**.

1. Decompose L such that $L = \Delta^T \Delta$

2: $H \leftarrow I - n^{-1} ee^T$

3: $\Psi \leftarrow \Delta^T HX$

4: **Compute the sparse basis based on the PMD method:**

Let $\Psi^1 \leftarrow \Psi$

For $k \in 1, ..., K$:

Find u_k, v_k and d_k by applying the following single-factor PMD algorithm to Ψ^k :

Initialize v_k to have L_2-norm equal to one.

Repeat (a) and (b) until convergence:

(a) $u_k = \frac{U_{K-1}^{\perp} U_{k-1}^{\perp T} \Psi^k v_k}{\left\| U_{k-1}^{\perp T} \Psi^k v_k \right\|_2}$

(b) $v_k = \frac{S(a,\tau)}{\|S(a,\tau)\|_2}$, where $a = \psi^k u_k$, $\tau = 0$ if $\|v_k\|_1 \leq c_2$ otherwise an appropriate τ is found so that, the condition is fulfilled.

$d_k \leftarrow u_k^T \Psi^k v_k$.

$\Psi^{k+1} \leftarrow \Psi^k - d_k u_k v_k^T$

5: **Encode training data:** $Z \leftarrow XV$

6: **Encode test data:** $z \leftarrow xV$

The regularization parameter, c_2 controls the sparsity of the Eigen vectors. Algorithm 4 shows the procedures for SSPCA. As can be seen, the K-factor $PMD(.,L_1)$ is utilized for finding the row and column vectors u_k and v_k respectively. The update equation for u_k forces orthogonality. $U_{k-1}^{\perp}$ is an orthogonal basis to $U_{k-1} = \{u_1, u_2, ..., u_{k-1}\}$. This update step yields orthogonal factors. It can not be used directly for v_k, since it does not result in a sparse solution. However, the v_ks are not very correlated, since they are associated with orthogonal u_ks (Witten et al., 2009). In the update equation for v_k , the S denotes the soft thresholding operator, so that for $\tau > 0$:

$$S(a,\tau) = \begin{cases} \text{sgn}(a)(|a| - \tau) & |a| > \tau, \\ 0 & |a| \leq \tau. \end{cases} \tag{C.11}$$

The solution to the above equation, satisfies $v_k = \frac{S(a,\tau)}{\|S(a,\tau)\|_2}$ with $\tau = 0$, if this results in $\|v_k\|_1 \leq c_2$; otherwise, τ is chosen so that $\|v_k\|_1 = c_2$. Further demonstrations for these update formula can be found in (Witten et al., 2009) and also provided in the Appendix.

In fact, the use of soft thresholding inside the convergence loop, reduces the absolute value of the Eigen vector elements so that, some of them will become zero or close to zero. The features that, the target vector Y is dependent on (relevant features), should remain among the non-zero elements and the zero or small elements should correspond to the irrelevant and noisy input variables. This should happen if the kernel and other parameters are chosen appropriately, specially for the first Eigen vector when the original Ψ^1 is used. An appropriate kernel is the one that has the highest dependency to the input matrix or in other words, is close to the target Y behavior. With an appropriate penalization or constraint value c_2, most irrelevant variables should be canceled out and most relevant ones should remain. Then, the result of such maximization is a sparse Eigen vector that the algorithm converges to. As a result, some rows of zero are formed in the final Eigen matrix corresponding to the common zero elements of the Eigen vectors.

Even with similar penalization values, the sparsity of the SSPCA and PMD-based SPCA are not necessarily similar. Because the objective function of SSPCA includes a kernel of the target vector $\Psi_{n\times p} = \Delta^T H X$, while the SPCA objective uses only the input matrix X. This can increase the sparsity level of the proposed method than SPCA. Because in computation of the Ψ based on Δ, the Eigen value matrix of the kernel is used. Depending on the rank of the kernel, it is likely that many diagonal elements of the Eigen value matrix become small or zero. Thus, the sparsity of Ψ will be increased.

C.4 Comparison with SPLS method

Due to the closeness of the proposed method to SPLS (Chun and Keles, 2010), their main differences is described here. SPLS is a sparse version of the well known supervised regression method PLS. In PLS, the response matrix $Y_{n\times q}$ and the predictor matrix $X_{n\times p}$ are decomposed into latent vectors so that, $Y = TQ^T + F$ and $X = TP^T + E$. $T_{n\times k}$ is a matrix that produces K linear combinations (scores), $P_{p\times k}$ and $Q_{q\times k}$ are matrices of coefficients (loadings) and $E_{n\times p}$ and $F_{n\times q}$ are matrices of random errors. PLS finds the columns of $W = (w_1, w_2, ..., w_K)$ by successive optimization problems and then, the latent component matrix $T = XW$ is computed:

$$w_k = \arg\max_w cor^2(Y, Xw)var(Xw) \quad \text{s.t.} w^T w = 1, \quad w^T \Sigma_{XX} w_j = 0 \ , \tag{C.12}$$

for $j = 1, ..., k-1$, where Σ_{XX} is the covariance of X. Using the statistically inspired modification of PLS (SIMPLS), the k^{th} estimated direction vector $\hat{w}_k$ is found by solving the following optimization problem:

$$\hat{w}_k = \arg\max_w \ w^T \sigma_{XY} \sigma_{XY} w \quad \text{s.t.} \quad w^T w = 1, \quad w^T \Sigma_{XX} w_j = 0 \ , \tag{C.13}$$

Σ_{XX} and σ_{XY} are the populations covariances of X and Y that can be replaced by the samples covariances (S_{XX}, S_{XY}):

$$w_k = \arg\max_w \ w^T X^T Y Y^T X w \quad \text{s.t.} \quad w^T w = 1, \quad w^T S_{XX} w_j = 0 \ . \tag{C.14}$$

Using W, the latent components T and loadings Q are computed. Finally, $\hat{\beta}_{PLS}$ is obtained by $\hat{\beta}_{PLS} = \hat{W}\hat{Q}^T$.

In the sparse version of the PLS algorithm, an L_1 penalty is imposed to the PLS objective function:

$$w_k = \arg\max_w \ w^T X^T Y Y^T X w \quad \text{s.t.} \quad w^T w = 1, \quad |w| \leq \lambda \tag{C.15}$$

.

This optimization problem is solved by a bi-convex procedure that is explained in more detail in (Chun and Keles, 2010).

The major difference between the SPLS and SSPCA can be explained by the definition of the correlation and dependency. Similar to PLS, SPLS aims to maximize the covariance between two random variables while SSPCA (similar to supervised PCA) maximizes the dependency between them. In other words, SPLS can detect linear dependence between two variables while in SSPCA any linear or non-linear dependency can be detected. This is performed by the choice of an appropriate kernel. In addition, after finding $\hat{\beta}_{SPLS}$, a linear regression is performed to compute $\hat{Y}$. However, SSPCA is a pre-processing step and can be followed by different regression or classification methods. The differences between supervised PCA and PLS are explained in more detail in (Barshan et al., 2011). They are the same for SPLS and SSPCA.

C.5 Experimental results

Five methods including PCA, SPCA based on the PMD method, supervised PCA, SSPCA and SPLS were applied on three simulated and three real data sets and the results will be shown in this section. Both regression and classification scenarios exist among these data sets. In data simulations, both linear and non-linear conditions were generated. In all the experiments and for all the methods, at least three Eigen vectors were chosen, so that their corresponding Eigen values explain at least 95% of variance. The models were trained using the cross validation (CV) model selection technique. In both regression and classification problems, the support vector machine (SVM) from the LibSVM toolbox (Chang and Lin, 2011) was used in training over the CV loops and the final tests. We have also employed CV loops for selection of the SVM parameters such as kernel type, spread parameter of radial basis function (RBF), degrees of the polynomial kernels etc. For each data set, based on its dimension, the appropriate number of folds was determined. Since in many real problems, the number of data points is less than the number of features, such condition was considered. For example, in cases where the number of samples was much less than the number of variables ($N \ll P$), larger number of folds (e.g. 10 folds) were used to avoid over-fitting.

No model parameters were required to be found for PCA. However, for all the other methods, CV loops were used for model selection; In SPCA, CV was used

for the choice of the restriction parameter c_2. As mentioned in section C.2.2, c_2 can be chosen in the range of $1 \leq c_2 \leq \sqrt{P}$. In supervised PCA, CV was used for the choice of the kernel type. The tested kernels were RBF, adaptive (RBF) (Zelnik-manor and Perona, 2004), quadratic and sinusoid kernels. For RBF and quadratic kernels, the spread parameter σ and degree parameters were respectively chosen based on iterations over a list of candidate values. For the proposed SSPCA method, both c_2 and kernel were found based on CV. The required parameters for SPLS such as λ are also found using a CV loop.

Root Mean Square Error (RMSE) was used as an evaluation criterion for all the methods in the regression problems for both the training (over the CV loops) and final tests. In the case of classification, the percentage of classification performance was considered. In addition, the average number of non-zero rows in the selected Eigen vectors are reported. All analyses were performed using MATLAB (R2013a).

C.5.1 Simulation results

The first sets of experiments were performed on some simulated data sets to evaluate the performance of the proposed SSPCA method and compare it with the other methods. In these experiments, the first Eigen vector will be plotted. This helps to compare the sparsity level of the tested algorithms as well as their ability to find the relevant features. As mentioned in section C.3, in the case of SSPCA, the Eigen vectors elements corresponding to the irrelevant and noisy variables should be zero or small in absolute values, while those corresponding to the relevant features should be higher in absolute values. Specially, when the kernel type and other parameters are chosen appropriately. Generally a successful method should have small (zero if it is an sparse method) elements for irrelevant and noisy variables and higher absolute values where the variables are relevant. That is, the principal directions should mostly be formed by the relevant features.

In all simulations, the data set was randomly divided into training and test sets 5 times and the average results were considered.

C.5.1.1 Simulation 1

In this example, a data matrix $X_{sim1(150\times120)}$ with $N = 150$ random samples and $P = 120$ variables were generated from a standard normal distribution.

Table C.1: Regression results for the first simulated data set.

	PCA	SPCA	Sup. PCA	SSPCA	SPLS
$RMSE_{tr}$	9.57±0.58	9.76±0.83	5.00±0.17	2.14±0.91	0.00±0.00
$RMSE_{ts}$	9.33±1.00	9.75±1.24	7.69±0.65	2.53±1.42	0.00±0.00
Num. of NZ.	120.00±0.00	43.00±17.46	120.00±0.00	12.8±5.40	10.00±1.00

Then, a linear function of four variables $X(5, 15, 25, 35)$ was defined:

$$Y_{sim1} = 6X_{sim1}(5) + 5X_{sim1}(15) - 7X_{sim1}(25) - 3X_{sim1}(35). \tag{C.16}$$

There were 100 training samples and 50 test samples. The training set was used for finding the Eigen vectors. Fig. C.1 shows the first Eigen vector for the PCA, SPCA, supervised PCA and SSPCA methods as well as the regression coefficients of SPLS (β_{SPLS}). In this example, β_{SPLS} was scaled to be shown on the same plot with the Eigen vectors. For ease of visualization,each method graph is plotted with an offset from other methods. The big circles with black edges show the relevant features. In the figure, the y axis shows the numerical value of Eigen vector elements. Based on its sign (positive or negative), each element is combined with others to form the principal direction for transforming data into the new space. Table C.1 shows the average regression results. The last row shows the average number of non-zero rows in the selected Eigen vectors. SPLS obtained the best result in terms of accuracy and sparsity and then the proposed method is the next best method for this linear function.

C.5.1.2 Simulation 2

The data matrix is $X_{sim2(100\times50)}$. The non-linear function depends on variables $X(10, 40)$:

$$Y_{sim2} = (1 + X_{sim2}(10)) \circ (1 + X_{sim2}(10)) + X_{sim2}(40) \oslash (0.5 + (1.5 + X_{sim2}(10)) \circ (1.5 + X_{sim2}(10))). \tag{C.17}$$

The $\circ$ and $\oslash$ show the element-wise multiplication and division respectively. The data set was divided 5 times randomly into training (30 samples) and test (70 samples) sets. Fig. C.2 and table C.2 show the results. Each method graph is plotted with an offset from others similar to the previous simulation. As can

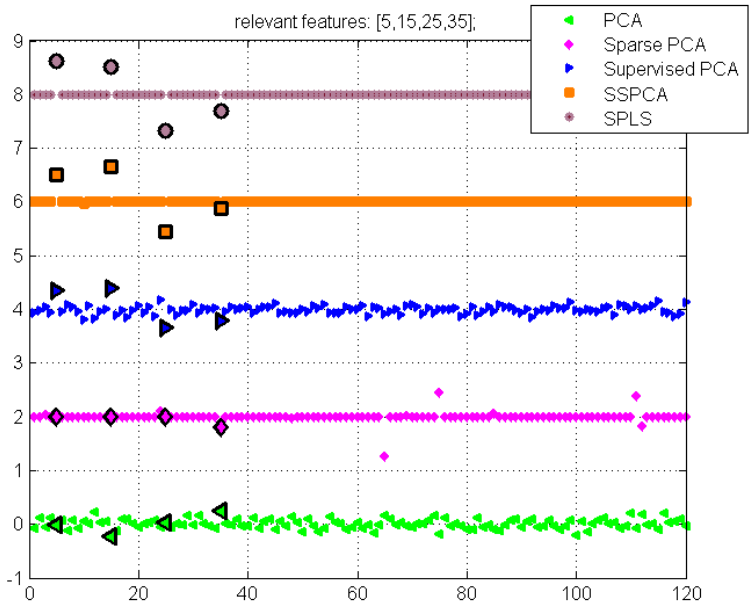


Figure C.1: Comparison of the first Eigen vector/regression coefficients of the five tested methods on the first simulated data set. The black edged circles show the relevant features.

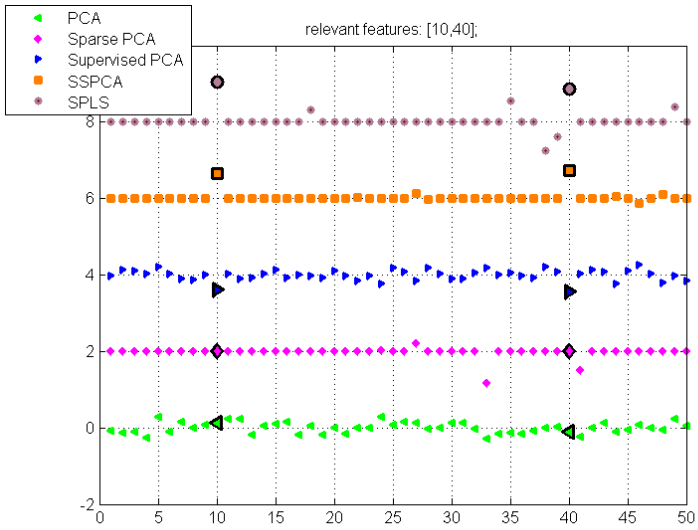


Figure C.2: Comparison of the first Eigen vector/regression coefficients of the five tested methods on the second simulated data set. The black edged circles show the relevant features.

Table C.2: Regression results for the second simulated data set.

	PCA	SPCA	Sup. PCA	SSPCA	SPLS
$RMSE_{tr}$	1.81±0.22	1.78±0.34	1.38±0.30	1.42±0.20	0.50±0.34
$RMSE_{ts}$	2.05±0.14	2.08±0.13	1.99±0.09	1.79±0.17	2.41±0.42
Num. of NZ.	50.00±0.00	10.80±1.30	50.00±0.00	13.40±4.39	22.60±16.62

be seen, for this non-linear function, SSPCA obtained the best result while the worst result was for the SPLS method. That is SPLS, as a linear regression method, is not an appropriate method for non-linear data sets.

C.5.1.3 Simulation 3

The data matrix is $X_{sim3(400\times30)}$. The non-linear function depends on variables $X(5, 20)$:

$$Y_{sim3} = exp(X_{sim3}(5)) - 2X_{sim3}(20) \circ X_{sim3}(20). \qquad \text{(C.18)}$$

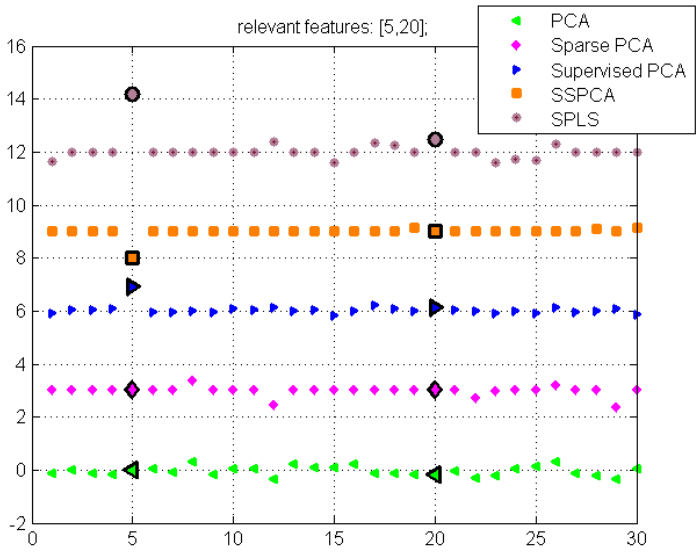


Figure C.3: Comparison of the first Eigen vector/regression coefficients of the five tested methods on the third simulated data set. Each method graph is shifted up with an offset for better visualization. The black edged circles show the relevant features.

The data set was divided 5 times randomly into training (300 samples) and test (100 samples) sets. Fig. C.3 and table C.3 show the results.

C.5.2 Real data sets results

In this part of the report, three real data sets are considered and the five methods are tested on them. In all cases, the data sets were divided 4 times into training and test sets and the average results are considered.

Table C.3: Regression results for the third simulated data set.

	PCA	SPCA	Sup. PCA	SSPCA	SPLS
$RMSE_{tr}$	3.55±0.44	3.40±0.54	2.54±0.30	2.59±0.25	3.02±0.35
$RMSE_{ts}$	3.61±1.16	3.51±1.13	2.85±0.70	2.75±0.75	3.36±0.97
Num. of NZ.	30.00±0.00	23.00±1.73	30.00±0.00	10.80±1.79	12.80±6.72

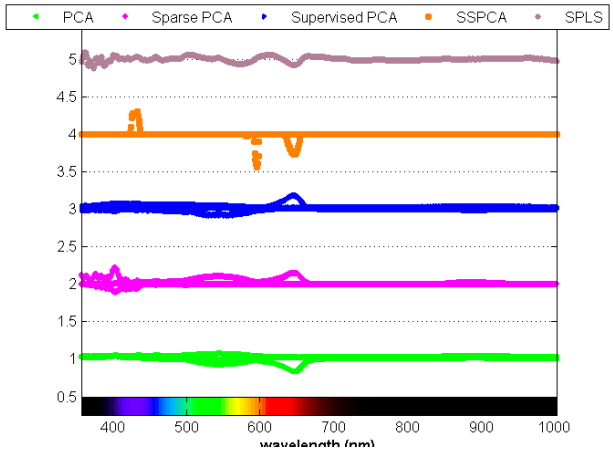


Figure C.4: Comparison of the first three Eigen vector/regression coefficients of the five tested methods on the apple data set.

C.5.2.1 Prediction of solvable solid content (SSC) of apple using spectroscopic measurements

The first real data set is the spectroscopic data of an apple type called *Rajka*. This is the same data set used in (Sharifzadeh et al., 2013a). Spectroscopic measurements were performed in 825 wavelengths (306 -1130 nm) and there were 185 data points (apple samples) in total. In addition, the SSC (%Brix) value for each apple was available from laboratory measurements. We divided the data into training and test sets 4 times based on a systematic sampling method called a smooth arrangement or smooth fractionator (Gundersen, 2002). For this aim, the samples were ranked in ascending order according to the SSC level. Then, from every 4 samples, one was chosen as test (unseen data during training) and the rest as training. By using this method, both training and test sets comprise the original variation of the data.

Fig. C.4 shows the first three Eigen vectors of the first four methods that are shown on the same plot together with the SPLS regression coefficients. The graphs are also shifted up in this plot similar to the previous illustrations. The average results are presented in table C.4. As can be seen, the proposed method is the best method in terms of accuracy and sparsity. SPLS and supervised PCA are the second best methods. However their number of used wavelengths are not comparable with the proposed method. All methods have a peak in the red color area of the visible bands that corresponds to the apple color.

Table C.4: Average regression results for the apple data set

	PCA	SPCA	Sup. PCA	SSPCA	SPLS
$RMSE_{tr}$	0.91±0.03	0.92±0.03	0.88±0.02	0.88±0.01	0.79±0.04
$RMSE_{ts}$	0.90±0.07	0.91±0.05	0.88±0.07	0.87±0.06	0.88±0.07
Num. of NZ.	825.00±0.00	439.75±177.86	825.00±0.00	149.00±202.38	778.25±48.93

Table C.5: Average regression results for the meat data set.

	PCA	SPCA	Sup. PCA	SSPCA	SPLS
$RMSE_{tr}$	2.32±0.09	2.42±0.51	2.25±0.36	1.93±0.15	1.06±0.07
$RMSE_{ts}$	2.32±0.22	2.36±0.72	2.52±0.14	2.01±0.32	1.60±0.23
Num. of NZ.	20.00±0.00	11.75±6.18	20.00±0.00	9.25±3.20	18.50±1.73

C.5.2.2 Prediction of a* color component for several meat types using multispectral images

This data set consists of multispectral images of different types of meat, e.g. turkey, chicken, beef, veal and pork. This data was previously used in (Sharifzadeh et al., 2014). Totally, there were spectral images in 20 wavelengths (430-970) and 52 meat samples. The median of the pixel values in an ROI was considered at each wavelength, forming a 52×20 matrix. In addition, the a* color component of each sample was available from a Minolta colorimeter measurement. The data was divided randomly into training and test sets 4 times. In each data set, the number of training and test samples were 38 and 14 respectively.

The first three Eigen vectors of the first four methods are shown in the same plot together with the regression coefficients of SPLS in Fig. C.5. β_{SPLS} is scaled in this plot. Here also, the graphs are visualized with an offset. The regression results are presented in table C.5. As can be seen, SPLS obtained the best result in terms of accuracy and SSPCA is the second most accurate method. However, SSPCA is the best method in terms of sparsity. SPLS uses most of the 20 wavelengths on average. Reducing the number of wavelengths is important for a vision system design in industrial scale. Both the red color wavelengths as well as the NIR bands are among the selected bands by the first three Eigen vectors of SSPCA. The red area corresponds to the color of most meat types and NIR regions are correlated to their chemical characteristics.

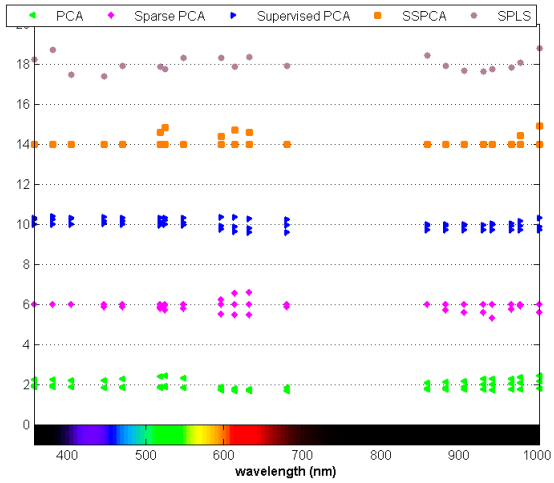


Figure C.5: Comparison of the first three Eigen vectors/regression coefficients of the five tested methods on the meat data.

Table C.6: Average regression results for the leukemia data set.

	PCA	SPCA	Sup. PCA	SSPCA	SPLS
PRF_{tr}	98.62±1.77	100.00±0.00	98.63±0.91	98.17±1.48	100.00±0.00
PRF_{ts}	97.30±3.13	94.44±7.86	94.52±7.86	94.52±4.54	95.91±5.30
Num. of NZ.	7129.00±0.00	2618.25±405.38	7129.00±0.00	30.75±18.34	1630.50±1671.96

C.5.2.3 Leukemia microarray classification and gene selection

The leukemia data set consist of 7129 genes and 72 samples (Goluband D. K. Slonim et al., 1999). Previously it was used in (Zou and Hastie, 2005). There are two types of leukemia (acute lymphoblastic leukemia and acute myeloid leukemia). The goal is to predict the type of leukemia based on the expression level of those 7219 genes. In microarray analysis, it is important to diagnose the related genes to the disease. In our experiment, we divided the data into training and test sets 4 times based on the smooth fractionator method (Gundersen, 2002), so that 75% of samples were chosen for training and the rest were kept for test. The percentages of classification performance as well as the number of selected genes are shown in table C.6. PCA obtained the best classification rate using all the genes while the other methods performances come close to that. However, in terms of gene selection, the proposed method obtained an excellent result compared to other methods.

C.6 Discussion

The experimental results from the simulations as well as the real data sets demonstrate that the proposed algorithm for SSPCA can make an appropriate trade off between the accuracy and sparsity. In the first simulation, SPLS was the best method in terms of accuracy and sparsity as there was a pure linear relationship between X and Y. This is due to the linear kernel in its objective function. However, in the case of non-linear relationships, the second and third simulation results showed that the SSPCA can perform better in terms of accuracy and sparsity. The choice of kernel type and penalization parameter play an important role on the accuracy and sparsity of this method. When the kernel is close to data behavior, the results can improve more. As expected, the sparsity of the SSPCA was better than SPCA in almost all cases and its accuracy was better than supervised PCA in all experiments due to canceling the effect of irrelevant and noisy variables.

Another important issue is that, SSPCA is both supervised and sparse. However, when in a data set the response is dependent to the maximum variation of the data, the supervision does not improve the result strongly compared to the unsupervised methods. This can explain the reason for the proposed method not achieving the best result in terms of accuracy compared to the other methods for some data sets.

SSPCA was also successful for high dimensional data sets such as the apple and microarray data.

Another important aspect of the SSPCA algorithm, is its ability on choosing the relevant features. This can be used as a criterion to perform feature selection as a pre-processing step for different applications.

C.7 Conclusion

In this paper, an SSPCA method was proposed for pre-processing of data sets with available target vectors. It computes sparse Eigen vectors based on the maximum dependency of the data to the response. The resulting Eigen vectors are almost orthogonal. The algorithm is based on the previous supervised PCA and penalizing terms were added to make the Eigen vectors sparse. The new objective function was solved based on the PMD algorithm. The SSPCA Eigen vectors are sparser compared to the PMD-based SPCA. Similar to the PMD method, the objective was solved as a bi-convex optimization problem. Due to

the use of the HSIC criterion in its objective function, this method can be used for data sets with linear as well as non-linear behavior. Experimental results showed that SSPCA can make an appropriate compromise between accuracy and sparsity. Comparison of the results from PCA, PMD-based SPCA, supervised PCA, SSPCA and SPLS on both simulated and real data sets showed that SSPCA works best in terms of sparsity. The accuracy was also comparable with the other methods. In addition, its sparse Eigen vector can be used as a means of feature selection, since the relevant features are usually among their non-zero elements.

Acknowledgment: This work was (in part) financed by the Center for Imaging Food Quality project which is funded by the Danish Council for Strategic Research (contract no 09-067039) within the Program Commission on Health, Food and Welfare.

Appendix

As mentioned in section C.2.2, SPCA can be solved as a $PMD(.,L_1)$ problem as it requires the column vector V to be sparse. In addition, in order to enforce orthogonality to the subsequent sparse PCs, in (Witten et al., 2009) an orthogonality constraint was added as follows:

$$\max_{u_k,v_k} \left\|u_k^T X^k v_k\right\| \textbf{ s.t. } \|v_k\|_1 \leq c_2, \|u_k\|_2^2 \leq 1, \|v_k\|_2^2 \leq 1, u_k \perp u_1, ..., u_{k-1} \tag{A.1}$$

With v_k fixed and $a = X^k v_k$, u_k is calculated based on the following steps:

$$\max_{u_k} \left\|u_k^T a\right\| \textbf{ s.t. } \|u_k\|_2^2 \leq 1, u_k \perp u_1, ..., u_{k-1} \tag{A.2}$$

Then $u_k = U_{k-1}^{\perp}\theta$, so that $U_{k-1}^{\perp}$ is an orthogonal basis to $U_{k-1} = \{u_1, u_2, ..., u_{k-1}\}$ and $\|u\|_2 = \|\theta\|_2$:

$$\max_{\theta} \theta^T U_{k-1}^{\perp T} X^k v_k, \textbf{s.t. } \|\theta\|_2^2 \leq 1, \tag{A.3}$$

The optimal θ is:

$$\theta_{opt.} = \frac{U_{k-1}^{\perp T} X^k v_k}{\left\|U_{k-1}^{\perp T} X^k v_k\right\|_2} \tag{A.4}$$

Therefore, the value for u_k is found:

$$u_k = \frac{U_{K-1}^{\perp} U_{k-1}^{\perp T} X^k v_k}{\left\|U_{k-1}^{\perp T} X^k v_k\right\|_2} = \frac{(I - U_{k-1}U_{k-1}^T) X^k v_k}{\left\|U_{k-1}^{\perp T} X^k v_k\right\|_2} \tag{A.5}$$

This update step is used for u_k in algorithm 4 and yields orthogonal factors.

With u_k fixed and $a = X^k u_k$,we have:

$$\max_{v_k} v_k^T a \quad \textbf{s.t.} \qquad \|v_k\|_2^2 \leq 1, \|v_k\|_1 \leq c_2 \tag{A.6}$$

or the equivalent minimization:

$$\min_{v_k} -v_k^T a \quad \textbf{s.t.} \quad \|v_k\|_2^2 \leq 1, \|v_k\|_1 \leq c_2 \tag{A.7}$$

The problem can be rewritten based on Lagrange multipliers:

$$-v_k^T a + \lambda \|v_k\|_2^2 + \tau \|v_k\|_1 \tag{A.8}$$

The Karush–Kuhn–Tucker conditions for optimality consist of :

$$\begin{array}{c} 0 = -a + 2\lambda v_k + \tau \Gamma_k \\ \lambda(\|v_k\|_2^2 - 1) = 0 \\ \tau(\|v_k\|_1 - c_2) = 0 \end{array} \tag{A.9}$$

where the first equation is obtained by differentiation and setting the derivative equal to 0. $\Gamma_k = \text{sgn}(v_k)$ if $v_k \neq 0$; otherwise, $\Gamma_k \in [-1, 1]$. If $\lambda > 0$, then from the first equation:

$$v_k = \frac{S(a, \tau)}{2\lambda} \tag{A.10}$$

In general, $\lambda = 0$ (if this results in a feasible solution) or it must be chosen such that $\|v_k\|_2 = 1$. Then as shown in algorithm 4:

$$v_k = \frac{S(a, \tau)}{\|S(a, \tau)\|_2} \tag{A.11}$$

Again by the Karush–Kuhn–Tucker conditions, $\tau = 0$ (if this results in a feasible solution) or it must be chosen such that $\|v_k\|_1 = c_2$. Then, $\tau = 0$ if this results in $\|v_k\|_1 \leq c_2$; otherwise, it is chosen such that $\|v_k\|_1 = c_2$.

Appendix D

An unsupervised feature selection strategy for characterization of VIS-NIR spectral signals of food products based on local maxima

Authors: Sara Sharifzadeh[1], Bjarne K. Ersbøll[1], Line H. Clemmensen[1].

1. Department of Applied Mathematics and Computer Science, Technical University of Denmark.

Technical Report

abstract

The use of spectral vision systems for quality monitoring of food items results in high dimensional signals. However, many of the wavelengths do not carry relevant information and might be highly correlated to each other, redundant or noisy. We introduce an unsupervised strategy that finds the appropriate features as a filter feature selection method. The proposed method uses the most significant peaks of the signal over the wavelengths for quality prediction. In order to avoid small local fluctuations on the signal envelop that result in identification of numerous peaks, a smoothing step is performed prior to the peak finding. This is useful especially, in cases that the input signal is noisy. In this paper, smoothing is performed based on adaptive thresholding of the wavelet coefficients. The proposed strategy is compared to the state of the art scale-space strategy based on Gaussian filtering which is a supervised method and also utilizes the significant local peaks of the signal. We also compare our work to two unsupervised feature selection strategies ; a filter solution based on an entropy function and a hybrid solution as a combination of a filtering step based on feature clustering followed by a wrapper frame work that uses FSSEM (Feature Subset Selection using Expectation-Maximization (EM) clustering). The results show that the proposed method is superior than the two other methods in terms of accuracy and is comparable to the supervised scale-space feature selection method. In terms of computation time, the proposed method is considerably faster than all other methods.

D.1 Introduction

Spectral vision systems have gained a lot of attention for food quality inspection. They also find application for medical purposes. Usually some quantitative values dependent to the acquired spectra should be predicted or classified. The spectra is obtained in high resolution and the spectral information are highly correlated. In addition, all of them are not relevant to the prediction or may be noisy. Therefore, feature selection should be performed to exclude the irrelevant and redundant features to reduce the complexity, dimensionality and over fitting problems.

Feature selection is part of dimension reduction strategies that can improve learning performance, lower computational complexity and build better generalizable models (Alelyani et al., 2013). Feature selection can be supervised or unsupervised. In this paper, unsupervised feature selection is considered which addresses the condition that training labels or target values are not available.

This can be the case in many real experiments where providing enough training samples or performing laboratory measurements are not possible.

Feature selection algorithms are categorized into filter, wrapper, hybrid, or embedded models. A filter model is independent of any classifier and each feature is evaluated by studying its characteristics using certain statistical criteria such as the Fisher score (Duda et al., 2000) or entropy function (Dash et al., 2002). A wrapper model utilizes a clustering algorithm to evaluate the quality of the selected features (Roth and Lange., 2003; Dy and Brodley, 2004). It starts by finding a subset of features and then the selected subset is evaluated based on a criterion such as likelihood or scatter separability for its clustering quality. These two steps are repeated until the desired quality is found. This method is accurate but computationally expensive. One important wrapper modeling is FSSEM (Dy and Brodley, 2000, 2004). The hybrid model employs a filter modeling and then it chooses the subset with the highest classification accuracy. Finally, an embedded model achieves model fitting and feature selection simultaneously (Zhao et al., 2010). More details about these models can be found in (Alelyani et al., 2013).

This report focuses on developing an unsupervised feature selection algorithm for high resolution spectral signals of food items. A previous work on unsupervised feature selection for spectroscopy data of food items was presented in (Krier et al., 2007). In that work, a methodology combining hierarchical constrained clustering of spectral variables and selection of clusters by mutual information was proposed. The mutual information measure was also used for mass spectrometry data [1] to find the relation between the features and the class labels in a supervised framework for detection of ovarian cancer through spectra of human serum (Krier et al., 2007). In (Prigent et al., 2010) classification of skin hyper-pigmentation was performed by spectral analysis of multi-spectral images. The spectrum data reduction was performed using projection pursuit.

In this report an unsupervised feature selection strategy is proposed based on the fact that, quality parameters of food items are related to their chemical composition or physical characteristics that influence their optical properties such as reflectance acquired by spectral measurements (Sun, 2009). As mentioned earlier the dimensionality of the spectral features are high and they are highly correlated. We hypothesize that the significant local peaks in the spectrum are

[1] A mass spectra is a plot of the ion signal as a function of the mass-to-charge ratio. The spectra are used to determine the elemental or isotopic signature of a sample, the masses of particles and of molecules, and to elucidate the chemical structures of molecules. Mass spectrometry works by ionizing chemical compounds to generate charged molecules and measuring their mass-to-charge ratios. It should not be confused with light spectroscopy. The type of spectra is totally different from spectral signals obtained from vision systems and instead of reflectance it shows the relative abundance of detected ions as a function of the mass-to-charge ratio.

related to the chemical or physical characteristics and can be used for prediction or classification of the quality parameters. Instead of all wavelengths only the local maxima are filtered and analyzed, such that, the algorithm should work faster compared to other selection methods that analyze all the features. In order to avoid small local fluctuations among the identified peaks, smoothing is performed prior to peak finding. This is important in cases where the spectra are noisy or the number of fluctuation on the envelope are considerable. This is performed based on adaptive thresholding of the wavelet coefficients of the spectra. Previously, a similar strategy used for variable noise suppression of the spectral data (Schlenke et al., 2012).

We have not seen any similar work based on local peaks for spectral data of food products. The use of local peaks for spectral information has been mostly used for Protein mass spectrometry data that is used for medical detection purposes. In a similar work (Tibshirani et al., 2004), classification of the protein mass spectrometry data for solid cancers was performed by peak probability contrasts. A list of all common peaks among the spectra was provided and their statistical significance and their relative importance in discriminating between the two groups of healthy and cancerous samples was tested in a supervised framework. In (Ceccarelli et al., 2009), feature selection and extraction was performed based on the theory of multi-scale spaces (Lindeberg, 1991) for high resolution mass spectrometry spectra.

To compare the proposed method with other techniques the scale-space feature selection strategy that is also based on local maxima is considered. This method was used in (Ceccarelli et al., 2009). In addition two other unsupervised feature selection methods were implemented; an unsupervised filter approach for feature selection (Dash et al., 2002) and the unsupervised FSSEM algorithm (Dy and Brodley, 2004) with an additional pre-selection step to reduce the computational cost.

The rest of this report is organized as follows. Section D.2 is about the materials and methods used in this paper. In section D.3 the experimental results are presented and we finalize with a discussion and a conclusion.

D.2 Materials and methods

In this section, we first describe the three methods used for comparison. Subsequently, we introduce the proposed unsupervised feature selection strategy. Finally, three spectral data sets of food items used for experiments will be described.

D.2.1 State of the art feature selection methods

This section reviews the three methods used for comparison to the proposed method.

D.2.1.1 Feature selection based on scale-space theory

Scale-space theory for signal analysis is a framework to find the local information of a signal (such as maxima or minima) when no prior information is available about them (Lindeberg, 1996). Therefore, the signal is represented at multiple scales to find the appropriate scales. In a multi-scale representation, structures at coarse scales constitute simplifications of corresponding structures at finer scales. In other words, the fine-scale information is successively suppressed as the scale increases. This principle preserves peaks or other features to be artificially introduced through scales and forces the analysis to be from finer scales to coarser scales (Ceccarelli et al., 2009). Thus, the peaks can give information about the spectrum.

In this paper a K-fold cross validation (CV) was used to train the models and the scale parameter of the standard deviation σ of a smoothing Gaussian kernel was varied inside the CV loop. At each CV iteration, for each scale parameter, the Gaussian kernel has applied and all the peaks, found for the training samples, are sorted. Then, the peaks were added to the model one by one (from the highest density to the lowest) and the error was computed at each step. Next, the minimum error was assigned to its corresponding number of peaks and scale in that CV iteration. At the end of all CV iterations, the validation error was averaged over all CV iterations and the scale and number of peaks corresponding to the minimum validation error was chosen for training and testing of the final models.

D.2.1.2 Unsupervised feature selection based on entropy function for clustering

In (Dash et al., 2002) a filter model criteria was used for feature selection that is based on the entropy-based distance. The main idea is that a proper subset of features must cluster data better than other subsets and a clustered data set has very different point to point distance histogram than data without clusters. The distance measure was used in computation of an entropy measure that assign low entropy to intra and inter-cluster distances, and a higher entropy to noisy distances. In other words, the relevant features with low entropy values

cluster data better than irrelevant features with high entropy values. Therefore, the best feature set was found based on the minimum entropy measures (Dash et al., 2002). The entropy measure based on the point to point distance is as follows:

$$E = -\sum_{X_i}\sum_{X_j} E_{ij} \tag{A.1}$$

$$E_{ij} = \begin{cases} \frac{\exp(\beta * D_{ij}) - \exp(0)}{\exp(\beta * \mu) - \exp(0)} & 0 \leq D_{ij} \leq \mu \\ \frac{\exp(\beta * (1.0 - D_{ij})) - \exp(0)}{\exp(\beta * (1.0 - \mu)) - \exp(0)} & \mu \leq D_{ij} \leq 1.0 \end{cases} \tag{A.2}$$

where D_{ij} is the normalized distance in the range $[0.0 - 1.0]$ between instances X_i and X_j and E_{ij} is normalized in the range $[0.0 - 1.0]$. β is a parameter that was set to 10 to assign sufficiently small entropy to intra- and inter cluster distances. As mentioned in (Dash et al., 2002), setting μ properly can help to distinguish between data with and data without clusters and it was computed using the equation A.2 by setting all other parameter based on the strategy explained in (Dash et al., 2002). This measure is able to assign a low entropy for data with clusters and a high entropy otherwise.

This method was used as an evaluation criteria for feature selection process. The other important step for this process is the search or generation step. In this search step for the best or optimal subset of features with minimum entropy, the forward selection algorithm was used (Dash et al., 2002).

D.2.1.3 Hybrid unsupervised feature selection based on FSSEM method

The FSSEM method has three different steps; feature search, clustering and feature subset selection criteria. In (Dy and Brodley, 2004), the sequential forward search (SFS) was used for feature search and the expectation maximization algorithm (EM) was used for clustering. Two different criteria were used for feature selection; Scatter separability criterion and maximum likelihood (ML). The number of clusters were found based on (Bouman, 1997). For initialization of the EM algorithm, the sub-sampling initialization algorithm proposed in (Fayyad et al., 1998) was used.

The scatter separability criterion finds features that best separate the data and the ML finds the features that model Gaussian clusters best (Dy and Brodley, 2004). In this work, the scatter separability criterion is used as the feature selection criteria .

Since the complexity of the FSSEM algorithm is high, the processing time for the high dimensional data sets used in this work is quite high and impractical. Therefore, a pre-selection of features is performed before applying this method in order to reduce the number of features and manage the computational time. For this aim, first the features with very low variance that remain almost unchanged over the samples are removed. Such features do not improve the discrimination or prediction. To reduce the dimensionality further, the features are partitioned into k clusters determined by the k components of the Gaussian mixture distribution. The clustering is iterated until the change in the likelihood function is negligible or a maximum number of iterations is reached. The number of mixture components is chosen to be high as it is not important to have clusters with overlap or close features. This is because, after this pre-selection step, the FSSEM algorithm will select the most relevant features among the pre-selected features in the next step.

D.2.2 The proposed method

As explained in section D.1, this work focuses on the unsupervised feature selection of the food spectral data. The high dimensional spectra of the food items represent the chemical composition and inherent physical properties of their constituent materials. From the illuminated energy (VIS-NIR), each material, reflect, scatter, absorb and/or emit the light in distinctive patterns at specific wavelengths that shows the spectral signature or fingerprint of that material. Therefore, the spectra contains more information than necessary (Michelsburg et al., 2012) and the relevant wavelengths showing the spectral signature are mostly located in the local peaks of the spectra which is well known from remote sensing and chemometric analysis (Serpico and Moser, 2007; Brereton, 2009). Figure D.1 shows the NIR spectroscopic signal of a case study from (Brereton, 2009) where samples of vegetable oils were assigned into one of the four classes (vegetable oil types) using pattern recognition techniques. As can be seen, discrimination between classes can be best performed at the local peaks.

Thus, unsupervised selection of the feature located on the local peaks of food items spectra can be performed for characterization, prediction or discrimination. In the first step, de-noising should be performed if there are local fluctuations on the envelope of the spectra to avoid finding lots of small peaks and smoothing the spectra. This is performed by thresholding the wavelet coefficients of the spectral signal. After that, the local peaks are found. A local peak is a wavelength of a sample that is larger in value than its two neighboring wavelengths. If a peak is flat, only one point (wavelength) among all the flat points is considered. Then, the density of peaks at each wavelength is calculated. This histogram is used to find the probability map of peaks for the wavelengths. Fi-

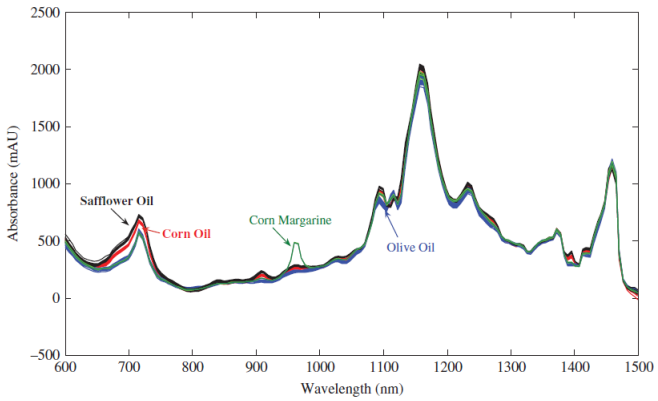


Figure D.1: NIR spectra of the four groups of oils (Brereton, 2009)

nally, by thresholding the probabilities, the peak wavelengths are found. There are different methods for finding the threshold such as the method proposed in (Kittler and Illingworth, 1986). In this work, we have chosen the threshold value simply at 0.5. As most of the spectral points of the data sets were non-peak points this threshold worked fine for the data sets, but it can be improved based on more accurate measurements for future studies.

D.2.2.1 Spectral data smoothing

Since spectral signals are often have some local fluctuations and might be corrupted by noise during their acquisition and transmission, smoothing methods should be employed on measurements in order to reduce such effects. In this work, the local fluctuations on the envelope of the spectra might result in finding lots of small peaks. Therefore, inspired by (Schlenke et al., 2012) a preprocessing step for smoothing the spectra and de-noising is performed based on thresholding of the wavelet coefficients of the spectral signal. The wavelet transform decomposes a signal into wavelet coefficients and small coefficients are linked to vibrations and noise effects whereas the large coefficients are related to significant features of the signal (Jansen, 2001). This is done using a base function called mother wavelet. Then, the small coefficients can be altered or removed by thresholding. If all the wavelet coefficients below a certain threshold are set to zero, it is called hard thresholding, while soft thresholding also reduces or shrinks the other wavelet coefficients that are higher than the threshold by

the chosen threshold as follows:

$$T_s(Y,t) = \begin{cases} \text{sign}(|\text{Y}| - \text{t}) & \text{if } |Y| \geq t \\ 0 & \text{else} \end{cases} \tag{A.3}$$

Another important point in this case is the choice of the optimal threshold. In practice, noise levels within a single real measurement are not necessarily constant and a constant threshold is not appropriate for all sections of the signal. Therefore, an ideal threshold is a variable function of the actual noise level.

In this work, the mother wavelet 'Symlets 8' is used and the soft thresholding together with an implemented MATLAB function for adaptive thresholding is employed. In this method, the threshold value is computed based on the changes in variance of noise in different time intervals (Lavielle, 1999).

D.2.3 Data description

Three different data sets are used in this report and are described in this section.

D.2.3.1 Spectroscopy measurements of apples (UV-VIS-NIR)

This data set was from an apple cultivar called Rajka. It was previously used in (Sharifzadeh et al., 2013a). Spectroscopic measurements were performed on both sides of apples, exposed and non-exposed to the sun, in 825 wavelengths (306-1130 nm) and the average results were considered. There were 185 data points (apple samples) in total. In addition, the solvable solid content (SSC) (%Brix) value for each apple was available from laboratory measurements. However, the SSC reference measurements were not used in the proposed feature selection method as it is an unsupervised method. They were just used in the evaluation step of the proposed method. Figure D.2 visualizes this spectroscopic data.

Similar to (Sharifzadeh et al., 2013a), we divided the data into training and test sets 4 times based on a systematic sampling method called a smooth arrangement or smooth fractionator (Gundersen, 2002). By using this method, both training and test sets comprise the original variation of the data. Each training set has 138 samples and each test set has 47 samples. Compared to the

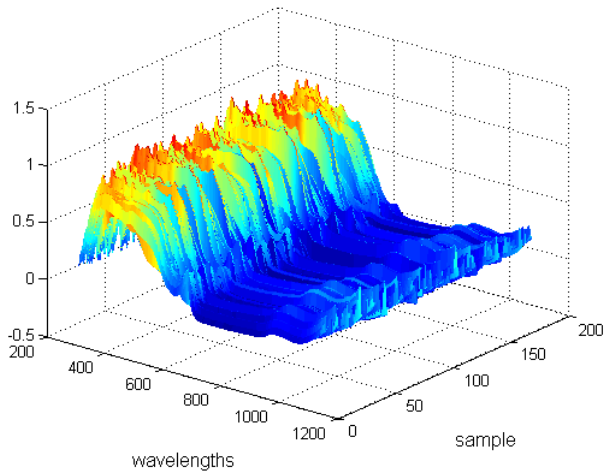


Figure D.2: Spectroscopy samples of apple type 'Rajka' in 825 wavelengths (VIS-NIR)

825 wavelengths, the number of samples are limited that makes the prediction difficult ($N \ll P$).

D.2.3.2 Hyper-spectral diffuse reflectance images of milk fermentation process (VIS-NIR)

This data set consisting of diffuse reflectance images of milk during fermentation process in the controlled condition for fat, temperature and protein factors. These images are obtained by illumination of a hyper-spectral coherent laser (480-1040 nm) into the surface of samples and a CCD camera captures the resulting profile. A complete description about this can be found in (Skytte et al., 2014). During the milk fermentation process, every 6 minute the hyper spectral imaging was performed in 57 wavelengths (480-1040 nm). This resulted in a spectro-temporal image set shown in the left side figure D.3. The process begins with a milk structure at t_1, and ends with a yogurt structure at t_{61}. The experiments were repeated for 8 times and in each round, the fat, protein and temperature level was controlled in low or high level, forming a total of 2^3 combinations. In addition, three experiments were conducted so that, all of the factors were in medium level.

Due to the high resolution of the acquired images a feature extraction strategy was used based on a slope parameter introduced in (Nielsen et al., 2011a,b). In

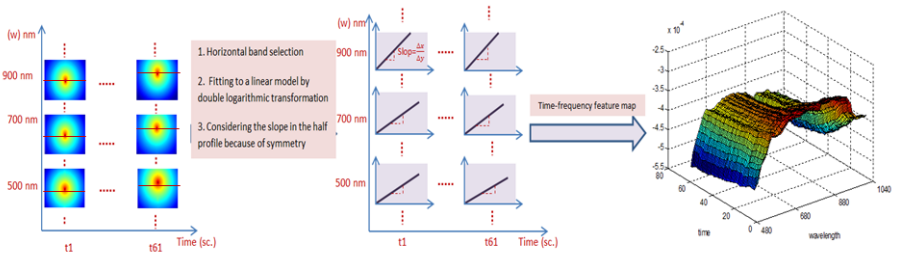


Figure D.3: The procedure of forming the feature map from a spectro-temporal image set of milk fermentation process; (left) The laser beam profile (Red corresponds to high pixel intensity and blue corresponds to low pixel intensity), (right) The slope features, (right) 3D representation of the final spectro-temporal feature map.

this method, from each image, a narrow band (11 pixels width) of the scattering profile from left to right including the scattering center is considered and averaged over the 11 pixels. The result is illustrated by a red line in the middle of each image in the left side figure D.3. The slope of the double logarithm transformation to the intensities along half of this line is as the final feature for each image. This is visualized in the middle of figure D.3. For more information in this case we refer to (Nielsen et al., 2011a,b). Finally, a 2D spectro-temporal feature map is formed from the slope values of the image sets. A 3D visualization of this map is shown in the right side of figure D.3. There are 57 elements along the wavelength and 61 along the time.

In this work, classification of the samples into one of the three levels of fat content using the spectro-temporal samples is addressed. In addition, reducing the number of wavelengths and time indexes is desired as this helps to simplify the vision set-up and the reduce the complexity of the practical experiments. This requires to perform feature selection along both wavelength and time. The 11 spectro-temporal data sets, were arranged two times to form two different pairs of sets $\{s_1, s_2\}$ and $\{s_3, s_4\}$. Each of the s_i $i = 1, ..., 4$ have two spectro-temporal samples with low and two with high level of fat contents. However, s_1 and s_3 have two samples with medium level of fat content and s_2 and s_4 have one. The two other controlled parameters (protein and temperature) are different in the samples. Then, each pair can be used two times for feature selection so that, for example, once s_1 is considered as the training set and s_2 as test and vice-versa. Therefore, experimental tests were applied four times on the prepared sets.

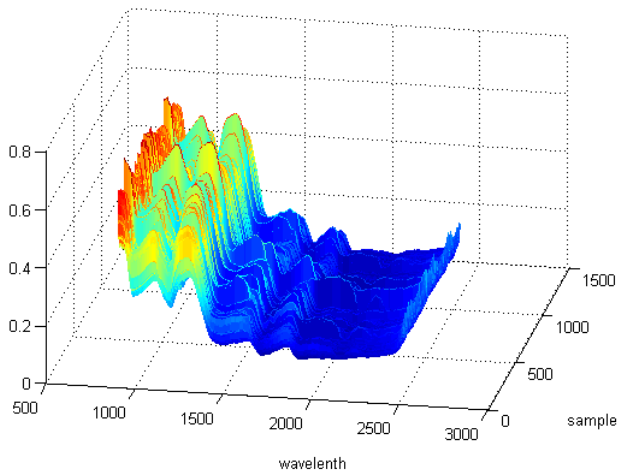


Figure D.4: The spectral data of 1042 fish pellets in 256 wavelengths

D.2.3.3 Hyper-spectral Images of aquaculture feed pellets (NIR)

The data set consists of hyper-spectral images of aquaculture feed pellets in the spectral range of 970-2500 nm in a step size of 6.3 nm, resulting in 256 spectral bands in the NIR range captured by a Specim vision system. The fill condition was used where there was white light in the background. The pellets used were coated with five different concentrations of added synthetic astaxanthin (0, 20, 40, 60, 80 ppm). This data set was used in (Ljungqvist et al., 2012). The aim of the study was to investigate the possibility of predicting the concentration level of synthetic astaxanthin coating of feed pellets by NIR hyper-spectral image analysis and to investigate what spectral features are of importance. The pellets were segmented from the background and divided into sub-regions of maximally 100×100 pixels and the mean of each region was used as a sample resulting in to 1042 samples ($N \gg P$). Figure D.4 shows the spectral samples of this data set.

In our work, we employed the smooth arrangement or smooth fractionator method (Gundersen, 2002) and divided the data based on this method into training and test sets four times.

D.2.4 Model evaluation

For each of the feature selection methods, the training data sets are used to find the best subset of features. The output of this step is the indexes of the best features among the input variables. In the next step, these indexes are used to selected the features of the test data. Finally the selected test features are evaluated. In order to evaluate the feature selection methods, prediction or classification is performed based on the type of data sets. For the spectroscopic apple data, prediction is performed using the support vector machine (SVM). For the spectro-temporal milk data set and spectral fish pellet data, classification based on SVM is employed. More information about the SVM can be found in many different sources such as (Hastie et al., 2009).

As an evaluation criterion, the average $RMSE$ for both training and test sets is estimated for regression problems. In the case of classification, the percentage of classification performance $PRF\%$ is considered.

D.3 Experimental Results

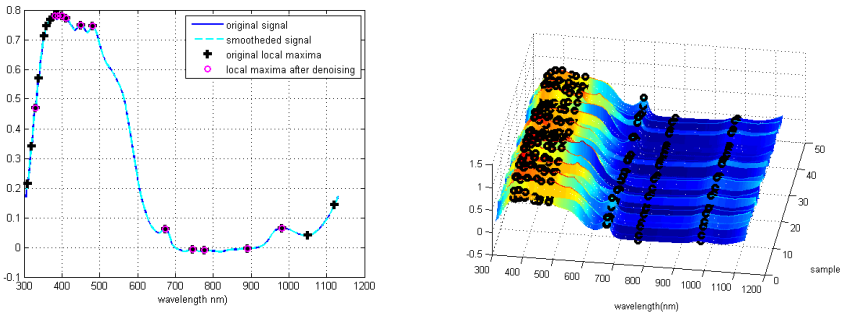
In this section, the results of the four methods, entropy filter, hybrid using clustering, scale-space and the proposed method are presented for the three data sets described in section D.2.3. The average and standard deviation of the results over the four training and test sets are reported.

D.3.1 Results of the apple spectroscopy data

Table D.1 shows the results of applying the four methods on the spectroscopic data of apples described in section D.2.3.1. As can be seen, the proposed method and the supervised scale-space method obtained better results compared to the other two unsupervised methods. The original as well as the de-noised signal for one sample are shown on the same plot in figure D.5-a. As can be seen, the number of peaks has reduced after smoothing. The final selected wavelengths are shown on the 3D illustration of this data set in figure D.5-b.

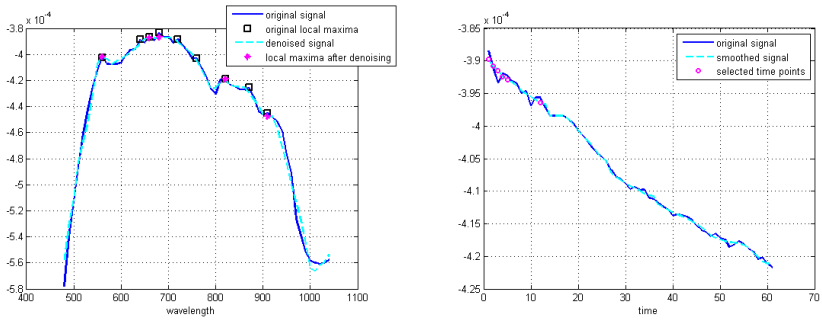
Table D.1: Comparison of the regression results for the apple data set

SVR	Apple data set	
	$RMSE_{tr}$	$RMSE_{ts}$
Entropy	1.09± 0.11	1.09± 0.07
Hybrid	1.04 ± 0.09	1.09± 0.07
Scale-space	0.87 ± 0.04	0.97± 0.05
Proposed method	0.91± 0.02	0.95± 0.07



(a) **(b)**

Figure D.5: The results of the proposed method on one sample of the apple data set (a) the original as well as the de-noised signal and their corresponding local maxima (b) the final selected peaks on the data set



(a) (b)

Figure D.6: Illustration of the proposed method on the milk data (a) the averaged spectro-temporal map over time (b) the averaged signal along the selected wavelengths.

D.3.2 Results of the spectro-temporal data of milk fermentation process

As explained in section D.2.3.2, there are only 11 2D spectro-temporal samples and the training and test sets have 6 (s_1, s_3) or 5 (s_2, s_4) samples. Due to the limited samples compared to the high number of variables (61 time points and 57 wavelengths), the feature selection is performed in two cascade steps; first the 2D profiles are averaged along the time (see figure D.6-a) and the best wavelengths are found by applying the feature selection strategies on the averaged 1-D profile. Then, averaging along the selected wavelengths is performed (see figure D.6-b) and feature selection results in finding the best time points. This reduces the number of variables. Since the over all 1D behavior along the time or wavelength in a 2D sample is almost consistent (see the right side of figure D.3) and only the height of the profile is different (e.g. between two different time points along the wavelength), the averaging strategy works for this data set. In other words, there is not a considerable shift in the location of peaks between the samples. The results of this data set are shown in table D.2. There might be some over fitting in the models due to the lack of samples in this data set.

D.3.3 Results of the hyper-spectral data of feed pellets

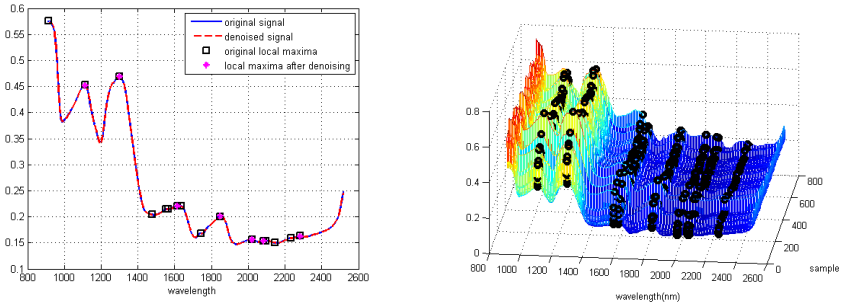
The results of the four methods on this data set are shown in table D.3. The proposed method as well as the scale-space gained better results again. The original as well as the de-noised signal for one sample are shown on the same

Table D.2: Comparison of the classification results for the milk data set

SVM	Milk	
	$PRF_{tr}\%$	$PRF_{ts}\%$
Entropy	100.0± 0.0	95.83 ± 8.33
Hybrid	100.0± 0.0	100.0± 0.0
Scale-space	100.0± 0.0	100.0± 0.0
Proposed method	100.0± 0.0	100.0± 0.0

Table D.3: Comparison of the classification results for the feed pellets data set

SVM	Fish pellet	
	$PRF_{tr}\%$	$PRF_{ts}\%$
Entropy	23.83± 5.18	21.49 ± 4.03
Hybrid	22.11 ± 2.28	22.27 ± 1.68
Scale-space	48.81 ± 18.69	49.63 ± 20.36
Proposed method	59.24 ± 2.44	56.63 ± 7.42



(a) (b)

Figure D.7: The results of the proposed method on one sample of the feed pellet data set (a) the original as well as the de-noised signal and their corresponding local maxima (b) the final selected peaks on the data set

plot in figure D.7-a. As can be seen, the number of peaks has reduced after smoothing. The final selected wavelengths are shown on the 3D illustration of this data set in figure D.7-b.

D.4 Discussion

Based on the results achieved by applying the four methods on the three data sets, the importance of considering the peaks in the analysis of the spectral data of food items is clear.

In figure D.8, the selected wavelengths by the four tested feature selection strategies are shown on the same plot. For ease of illustration, the spectral curves are shifted vertically. As can be seen, the scale-space method has considered most of the local peak points and after that the proposed method has used the most significant peaks. These two methods were more successful than the other methods according to the results obtained in previous section. Both of the entropy and hybrid methods use forward feature selection. The former uses the entropy criterion for feature selection. However, the group of selected features for which, the samples have the minimum entropy, are not located necessarily on the peaks. While the peak points have good correlation to the quality parameter (SSC). In the case of the hybrid method, the choice of features with highest variance at the primary step, keeps the peak points as there are lots of variations around the peaks. The forward selection step has also chosen some of the local peaks based on the scatter separability criterion.

Since the entropy based filtering and the hybrid methods are both based on forward selection algorithm, they require to examine all of the features that increases the complexity of these unsupervised algorithms. In addition, the scale-space strategy is a supervised frame work that uses k-fold CV loop with two internal loops for the choice of σ and the number of peaks. This also considerably increases the computational time. However, the proposed method does not use any of the above mentioned factors and hence is faster than the other methods. Table D.4 presents a comparison of the average computational times and the corresponding standard deviation of the tested methods for the apple data set. As can be seen, the scale space method has the highest computational time while for the proposed method it is considerably lower than the other methods.

Since the proposed feature selection method is unsupervised it can be used for the analysis of the quality of food items in the absence of the quantitative reference for quality in conditions that only spectral measurements are available. The short time of process makes it useful to be used in-line in real time projects.

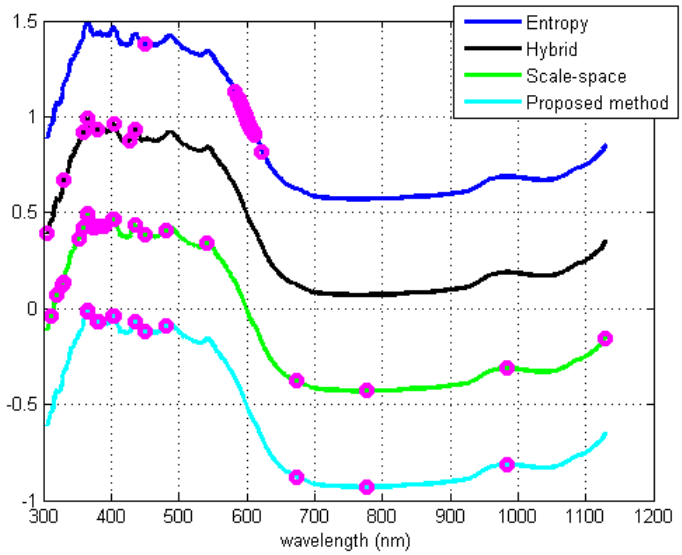


Figure D.8: Comparison of the selected wavelengths by the four tested methods for apple data set.

Table D.4: Comparison of the computational time of the tested methods

	Entropy	Hybrid	Scale-space	Proposed method
$t(sc.)$	10702.25±5621.40	7760.64±3448.79	36153.71±1853.81	122.37±83.05

In this work, the threshold value for finding the significant peaks was constant and further studies may be required for finding a more robust way for the choice of threshold value.

D.5 Conclusion

In this report, a new unsupervised feature selection method is proposed for spectral data of food products. The local extrema of the food items spectra are found for all training samples. Then, a histogram map for the peaks of all wavelengths is formed and thresholded to find the most important peaks. To remove the local fluctuations and noise effects in the spectral data, a smoothing step based on thresholding of the wavelet coefficients of the spectra is performed. The proposed method is compared with two other unsupervised feature selection algorithms; a filter method that uses an entropy function and a hybrid method based on clustering. In addition, the proposed method is compared with the state of the art scale-space strategy that is a supervised method. Experimental results showed that the proposed method is better than the filter and hybrid methods in terms of accuracy and comparable with the scale space method. It worked even better than the scale space method in some cases. In addition, it is superior than all other methods in terms of computational time. The proposed method is suitable for quality assessment of the food items using their spectral data in condition that no quality references are available.

Acknowledgment: This work was (in part) financed by the Center for Imaging Food Quality project which is funded by the Danish Council for Strategic Research (contract no 09-067039) within the Program Commission on Health, Food and Welfare.

Appendix E

A sampling approach for predicting the eating quality of apples using visible–near infrared spectroscopy

Authors: Mabel V Martíinez Vega[1], Sara Sharifzadeh[2],Dvoralai Wulfsohn[3], Thomas Skov [4], Line H. Clemmensen [2] and Torben B Toldam-Andersen[1].

1. Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen.
2. Department of Applied Mathematics and Computer Science, Technical University of Denmark.
3. Quality and Technology, Department of Food Sciences, Faculty of Science, University of Copenhagen.
4. Dayenú Ltda, San Fernando, Chile.

Published in Journal of the *Science of Food and Agriculture*.

Research Article

Received: 8 December 2012 Revised: XX XXXX Accepted article published: 30 April 2013 Published online in Wiley Online Library: 7 June 2013

(wileyonlinelibrary.com) DOI 10.1002/jsfa.6207

A sampling approach for predicting the eating quality of apples using visible–near infrared spectroscopy

Mabel V Martínez Vega,[a]* Sara Sharifzadeh,[b] Dvoralai Wulfsohn,[c] Thomas Skov,[d] Line Harder Clemmensen[b] and Torben B Toldam-Andersen[a]

Abstract

BACKGROUND: Visible–near infrared spectroscopy remains a method of increasing interest as a fast alternative for the evaluation of fruit quality. The success of the method is assumed to be achieved by using large sets of samples to produce robust calibration models. In this study we used representative samples of an early and a late season apple cultivar to evaluate model robustness (in terms of prediction ability and error) on the soluble solids content (SSC) and acidity prediction, in the wavelength range 400–1100 nm.

RESULTS: A total of 196 middle–early season and 219 late season apples (*Malus domestica* Borkh.) cvs 'Aroma' and 'Holsteiner Cox' samples were used to construct spectral models for SSC and acidity. Partial least squares (PLS), ridge regression (RR) and elastic net (EN) models were used to build prediction models. Furthermore, we compared three sub-sample arrangements for forming training and test sets ('smooth fractionator', by date of measurement after harvest and random). Using the 'smooth fractionator' sampling method, fewer spectral bands (26) and elastic net resulted in improved performance for SSC models of 'Aroma' apples, with a coefficient of variation $CV_{SSC} = 13\%$. The model showed consistently low errors and bias (PLS/EN: $R^2_{cal} = 0.60/0.60$; SEC = 0.88/0.88°Brix; $Bias_{cal} = 0.00/0.00$; $R^2_{val} = 0.33/0.44$; SEP = 1.14/1.03; $Bias_{val} = 0.04/0.03$). However, the prediction acidity and for SSC (CV = 5%) of the late cultivar 'Holsteiner Cox' produced inferior results as compared with 'Aroma'.

CONCLUSION: It was possible to construct local SSC and acidity calibration models for early season apple cultivars with CVs of SSC and acidity around 10%. The overall model performance of these data sets also depend on the proper selection of training and test sets. The 'smooth fractionator' protocol provided an objective method for obtaining training and test sets that capture the existing variability of the fruit samples for construction of visible–NIR prediction models. The implication is that by using such 'efficient' sampling methods for obtaining an initial sample of fruit that represents the variability of the population and for sub-sampling to form training and test sets it should be possible to use relatively small sample sizes to develop spectral predictions of fruit quality. Using feature selection and elastic net appears to improve the SSC model performance in terms of R^2, RMSECV and RMSEP for 'Aroma' apples.

Keywords: *Malus domestica*; SSC; representative sample; training set formation; variability

INTRODUCTION

The use of visible and near infrared spectroscopy (visible–NIR) for the rapid evaluation of fruit quality remains a topic of importance and interest for the food research community and food industry because, in a near future, it might be included in 'the tool box' for efficient farm management.[1,2] Spectral regions on the visible and near infrared spectrum have been used to predict quality in intact fruits such as apples (380 up to 2000 nm), apricots (600–2500 nm), citrus (636–1236 nm), grapes (650–1100 nm), kiwifruits (300–1100 nm), pineapples (400–2500 nm) with different degrees of success.[3] The fruit quality parameters studied with spectroscopy included: soluble solids content (SSC), firmness, acidity, dry matter, taste and starch, among others.[3] In most of these studies the quality characteristics were predicted using multivariate statistical models.

Two of the most important fruit quality traits are SSC and acidity.[4] These traits have a great influence on consumer liking and repetitive purchases. During fruit growth, the internal quality traits are expected to vary due to different causes (type of soil, weather, training and thinning techniques, etc.). This variation in quality might be the most important factor affecting the

* *Correspondence to: Mabel V Martínez Vega, Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen, Højbakkegård Allé 13, 2630 Taastrup, Denmark. E-mail: mmar@life.ku.dk*

a *Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen, Højbakkegård Allé 13, 2630 Taastrup, Denmark*

b *DTU Data Analysis, The Technical University of Denmark, Richard Petersens Plads, build. 305/123, 2800 Lyngby, Denmark*

c *Dayenú Ltda, San Fernando, Chile*

d *Quality and Technology, Department of Food Sciences, Faculty of Science, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C., Denmark*

calibration models, which are used to train different spectroscopy devices.[3] Model validation, an essential step to be carried out after calibration, has often been performed using samples from the same batch. The tendency has been to use or suggest large sets of samples, which together with pre-processing statistical methods, reached somewhat satisfactory results.[3–5] One conclusion was that the samples should be 'rich' in variation and ideally contain information from multiple orchards/seasons/cultivars to obtain sufficient robustness.[3,6–8] In addition, for the purpose of proper model construction, post-harvest sample arrangements have also been proposed with different aims. Interestingly, most of the studies reporting spectral robustness issues for fruit quality, used samples obtained randomly from either fruit trees or from the commercial market. Frequently, little has been reported regarding the sampling techniques applied during fruit collection and often relevant sample statistics (mean, standard deviation, ranges of the quality parameter of interest) have not been provided. As a result, the reproduction, comparison, evaluation and improvement of the mentioned experiments becomes challenging.

In an earlier study, we explored the variability of mass, sugar, firmness and starch of representative samples of 'Granny Smith' apples obtained at the orchard scale (Martínez Vega MV *et al.*, unpublished). In this study, we extend our approach of using the 'fractionator' tree sampling procedure to obtain representative apple fruit samples at time of harvest.[9] These samples were used to evaluate the performance of visible–NIR spectroscopy method for calibration and validation model development. Thus, the objectives of the study were: (1) evaluate the SSC and acidity prediction performance of an early and late season apple cultivar; and (2) to compare different sub-sampling techniques to form training and test sets on the overall performance of the prediction models. Furthermore, we discuss the main implications of the method in practice.

MATERIALS AND METHODS

Fruit material

Two Danish apple (*Malus domestica* Borkh.) cultivars, an early season 'Aroma' and a late season 'Holsteiner Cox', were collected at fruit maturity, in September and October 2011, respectively from 11-year-old trees at the Pometum orchard, University of Copenhagen, Denmark. The samples were selected using the 'fractionator' procedure for trees,[9] from 10 trees per cultivar. The fractionator procedure is a form of multi-level systematic uniform random cluster sampling, in which the trees, primary branches, and, at the final stage, branch segments form the clusters of fruit for sampling purposes. For both cultivars we used systematic sampling periods of 2 (for branch) and 2 for in-branch segment with random starts. When the branch segments bore more than one fruit, a random number was used to select one fruit from each of the final sample of branch segments. Each sampled fruit was labelled with a number to preserve information about the picking order. Once harvested, the samples were kept at room temperature (18 °C). To widen the spread of fruit SSC and acidity values for the experiments, apple quality measurements for each cultivar were performed after 5 (Date 1) and 10 (Date 2) days of storage. Likewise, to preserve the distribution of fruits per tree from the original sample, the sub-groups for each of the storage periods mentioned were selected by taking a systematic sample of fruit with period 2 while preserving the original picking order from the 10 trees. Thus, sample Date 1 contained apples 1, 3, 5 . . . , and Date 2 sample consisted of the fruit ranked 2, 4, 6 and so forth.

Determination of fruit quality

On each apple, two pieces of fruit flesh (stem to calix end), from the exposed and non-exposed side of the fruit were squeezed. Its juice was presented to a calibrated handheld brix meter (Metler Toledo 'Quick brix 60'; Mettler Toledo Inc. Columbus, Ohio, USA) to measure SSC content. The remaining juice was kept for acidity determination. Acidity was measured with a titrino (719 S Titrino Metrohm; , Herisau, Switzerland). The titration consisted of adding a solution of NaOH of concentration 0.1 mol L^{-1} to 5 mL a sample solution of apple juice until the pH reached 8.1. Results were expressed in grams of malic acid (the most abundant acid in apples) per 100 mL of apple juice.

Spectral measurements

A spectrometer (MOE-1 System, Tec5 AG, Oberursel, Germany) with MMS sensors and a 12 V/100 W halogen lamp was used to collect reflectance readings in 1 nm increments within a wavelength range between 400–1130 nm, yielding 731 values per spectrum. A calibration was performed using a white piece of barium sulfate every 20 apples. Spectral measurements were performed on the exposed and non-exposed (to sun) parts around the equator of each apple. A distance between the lamp and the fruit of 10 mm was maintained. A holder supported fruits to direct light in a 45° angle to avoid specular reflectance. The integration time was 161 ms. Each intact fruit was placed on a rotary circular base with the stem–calyx vertical and four equidistant guides on the base made sure that the measurements were approximately equidistant. The scans collected at each sample point were averaged and transformed to absorbance [log(1/reflectance)].[10]

Training and test sets arrangements

First, over-mature or damaged fruit samples on 'Date 2' were removed from the data sets. Then, three different data sets were formed for each cultivar.

Set A

A smooth arrangement from all samples ('Date 1' and 'Date 2' together) according to SSC and acidity values was performed. The 'smooth' arrangement was formed by ranking all the original sample of fruit in increasing order according to the SSC or acidity level, respectively, for SSC and acidity modelling. Then every second fruit was pushed out to form a monotonically increasing and then decreasing ordering of fruit by quality. From this new ordering, a predefined systematic sampling interval of '4' (probability $p = 1/4$) was applied to obtain approximately 25% of the samples for the test set. The remaining 75% of the samples comprised the training set. This procedure was repeated four times, starting with fruit ranked 1, 2, 3 and 4, corresponding to the four possible 'random starts' that form all possible systematic samples from the 'smooth' ordering. Systematic sampling from a smooth arrangement ('smooth fractionator') is a procedure designed to provide samples with high within-sample variance and low between-sample variance, which in this case means that both training and test sets capture well the variation of SSC and of acidity existing in the original sample.[11,12] The averages of results were used to evaluate the general performance of the regression methods on the data sets.

Set B

SSC samples of each cultivar from 'Date 1' formed the training set and samples from 'Date 2' the test set. The same criterion was used to construct the models for acidity.

Set C

The original data set was divided into training (75%) and test (25%) sets using simple random sampling without replacement. This was repeated 25 times to obtain 25 independent sets for training and testing. The averages of results were used to evaluate the general performance of the regression methods on the data sets.

Preprocessing of spectral data

Since the spectral data contained NIR bands, Multiplicative scatter correction (MSC) was applied.[10] In addition, because of the presence of visual bands, the original data set without MSC was also considered. All the models and algorithms were calculated using Matlab software (version R2011a; The MathWorks Inc., Natick, MA, USA).

Calculation of calibration and prediction models

Three different linear regression methods were used on each data set. For all the regression methods, 10-fold cross validation with a modified version of the standard error rule[13] was used for finding the best parameters to train the model.

Partial least squares regression

The commonly used partial least squares regression (PLS) method was used to predict fruit quality from spectra data. The basis of the method is to link the variation in the spectral information to the response to find only the relevant information for predicting the response.[10]

Calibration and prediction models were constructed using the 'internal validation' approach (using samples from the same batch).[3] The SSC and acidity data were autoscaled before model calculation. This latter procedure ensures that all samples have approximately the same contribution to the model.[10]

Ridge regression

This method is based on the penalisation of the regression coefficients. As a result, the regression model is regularised to reduce the variance of the predicted output.[13] The purpose is to alleviate the effect of noise on the model. Ridge regression requires that both the response vector (Y) and the data matrix (X) to be centred.

Elastic net

Elastic net (EN) is a sparse regression method based on the regularisation of regression coefficients. This means that the regression coefficients are shrunk so that some of them are set to zero. Therefore, it can cancel out the noise effect. In addition, it has a grouping effect and the non-zero coefficients correspond to the groups of correlated variables (wavelengths). When the number of variables (e.g. number of spectral bands = 731) is higher than the number of observations (e.g. $N_{SSC} = 196$ data points), the prediction becomes an 'ill-posed' problem[13] and EN is one of the appropriate methods in this case. This method requires the response vector (Y) to be centred and the data matrix (X) to be normalised with unique length for each variable.[14]

Feature selection

This method is commonly used for high dimensional data to reduce the complexity of the model. Since the dimensionality of the apple data was high (731 spectral bands), this pre-processing step was also employed. It was compared with the regression results using all the features (wavelengths). Feature selection helps to distinguish the wavelengths that carry the useful information for the prediction to simplify the model.

A common approach for dimension reduction is principal component analysis (PCA), but it is not an appropriate method for 'ill-posed' problems.[15] Although PCA is a dimension reduction method, each principal component is a linear combination of all the basic features (wavelengths). This means that it could not be used as a tool for reducing the number of used wavelengths for prediction. We applied a feature selection algorithm proposed in a former study (Sharifzadeh S *et al.*,[16] unpublished). The method first sorts the wavelengths according to the number of times that their corresponding regression coefficients were non-zero in several iterations of elastic net regression and then selects a subset of them as described below.

For this research, the regression coefficients obtained from applying EN on the set C (25 randomly generated training sets), were used for feature selection. First, the number of times that the coefficients were non-zero in each band was counted ('frequency of being non-zero'). Then, the wavelengths were sorted according to their corresponding frequencies. To choose a proper number of wavelengths for performing the regression task, a candidate list of the number of selected wavelengths was formed:

$$\text{candidate list of top selected wavelengths} = [20, 50, 80, 100, 150, 200, 250, 700]$$

In the next step, an EN regression with 10-fold cross validation was applied on only the 25 training sets using the spectral data corresponding to each of these candidate numbers of wavelengths. Finally, the best candidate number of wavelengths was chosen according to the corresponding minimum root mean square error of prediction (RMSEP).

Model evaluation

Model robustness was evaluated in terms of the coefficient of determination (R^2), the standard deviation (SD) of training and test sets, the standard error of calibration (SEC), the root mean square of residual errors of cross validation (RMSECV), the standard error of prediction (SEP), the root mean square of residual errors of prediction (RMSEP) and the bias.

RESULTS AND DISCUSSION

Cultivar variability along the harvest season

The 'fractionator' procedure yielded in total 205 fruits for 'Aroma' and 221 fruits for 'Holsteiner Cox'. The total number of samples for 'Aroma' in Date 1 was $N = 103$ fruit and for Date 2 was $N = 102$ fruit. The number of samples for 'Holsteiner Cox' was $N = 111$ fruit in Date 1 and $N = 110$ fruit in Date 2. Figure 1 illustrates the spread of SSC and acidity values for both cultivars after elimination of damaged samples (over-mature or with disease). The higher SSC values of 'Holsteiner Cox' were expected given the reported sweetness properties of the late season cultivar as compared to the early season 'Aroma'.[17]

In general SSC and acidity values had low to moderate variation. 'Aroma' samples had the same average of SSC on both post-storage dates, whereas 'Holsteiner Cox' samples showed a slight increase of the average SSC values. The increase is related to degradation of starch which normally is present at high levels in late season

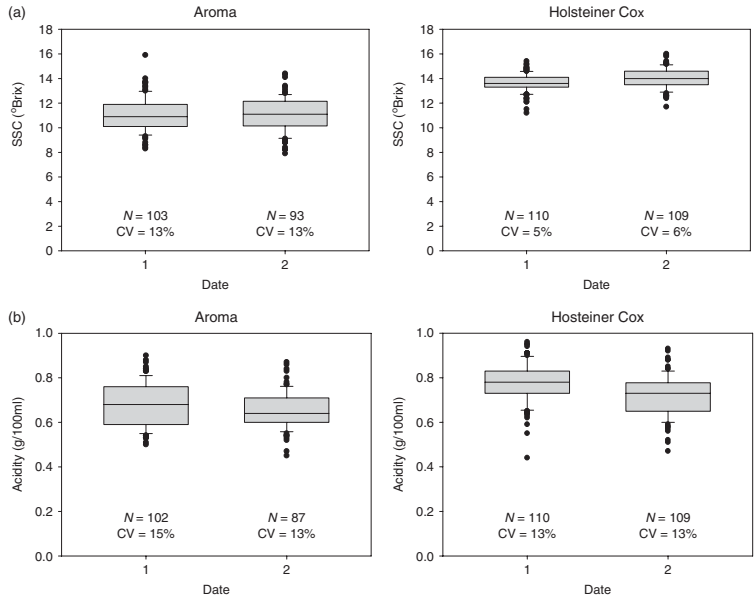


Figure 1. Box and whisker plots for (a) SSC (soluble solids content) and (b) acidity values for cultivars 'Aroma' and 'Holsteiner Cox' on two post-storage measurement dates (5 and 10 days). Extreme values, present for both variables and measurement dates are indicated by solid symbols on the plots. *N* = number of samples; CV = SD/mean.

cultivars with potential for postharvest storage (unpublished data). The lower coefficient of variation (CV = SD/mean) of SSC for 'Holsteiner Cox' as compared to 'Aroma' , showed that 'Holsteiner Cox' samples had a notably narrower spread of SSC values (Fig. 1a). 'Aroma' and 'Holsteiner Cox' had almost similar CVs of acidity on both harvest dates.

Spectral signatures of the early and late season cultivar

The spectral signatures for the apple cultivars 'Aroma' and 'Holsteiner Cox' obtained in two different post-storage dates are illustrated in Fig. 2.

There were differences in the shapes of the spectral signature between cultivars and between measurement dates. Furthermore, the curves showed large variability in absorbance at a given wavelength. The visible region (below 700 nm) of the spectra appeared more irregular than the NIR region (above 700 nm) between Dates 1 and 2.

The 'Aroma' signature showed a noisy area in the blue region 400–500 nm. On the further green region 500–600 nm, there were differences on the turning points of the curve between Date 1 and Date 2 spectra. Different spectral regions have been related to chemical components such as chlorophyll at 650–695 nm[18] or carotenoids and anthocyanins at shorter wavelengths than 650 nm, sugars in 470–484 nm, 498–512 nm, 526–540 nm, 568–582 nm, 665–679 nm,[19] and sour taste (acidity) in the 640–700 nm region.[20] The low absorption values around the area between 700 and 900 nm probably do not contain important information for 'Aroma' and 'Holsteiner Cox' cultivars. In this sense, spectral bands with almost zero light absorption have been reported to be influenced mainly by scattering properties of the tissue.[21] These spectral regions were not removed for the model calculations, however.

The peaked-shaped area shape above 900 nm is consistent with previous studies of SSC on apples. One should expect to find spectral curve peaks at around 800 nm related also to chlorophyll content,[22] 950 nm peaked areas may be related to water content and sugar–water peaks at 840 and 890 nm[23] and the overtones of the hydroxy (O—H) stretch/vibration of H_2O/carbohydrates may be explained at 930–1080 nm as well as variations in the absorption at 960 and 1060 nm, which are related to absorption of pure water and solutions of different sugar concentrations.[18]

Results of band selection

Figure 3 shows the counts of non-zero coefficients for each of the 731 wavelengths. The plot corresponds to the analysis of the original SSC data without MSC pre-treatment.

Figure 4 shows the resulting averages of the RMSEP values plotted after EN was applied on bands of training data according to the candidate list. As described previously in the feature selection section.

For SSC, the minimum RMSEP occurred at 600 features, but there was a very close RMSEP value also at 350. Because 350 bands was considerably smaller than 600, the first 350 top bands were selected for SSC. The same procedure was performed for acidity. In this case, the first 250 bands were chosen. The selected bands for SSC and acidity are shown in Fig. 5.

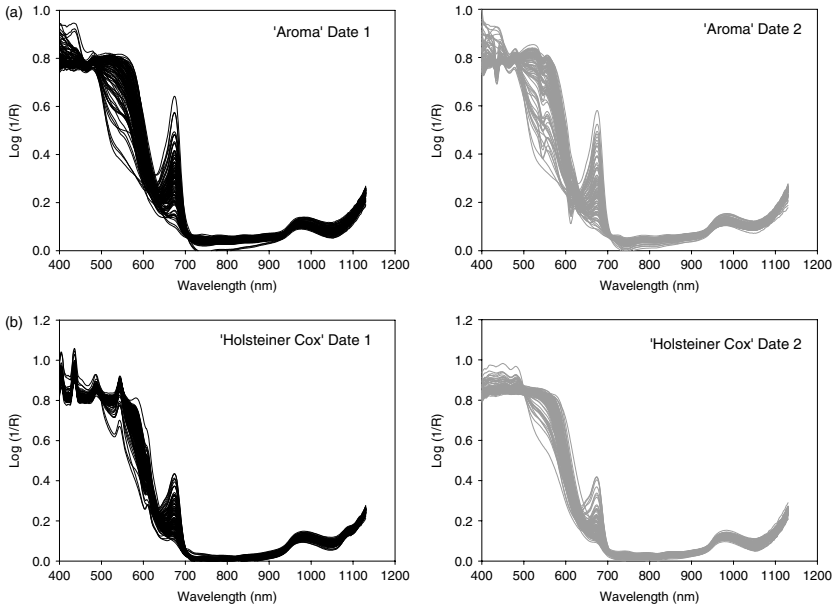


Figure 2. Raw spectral patterns recorded in the visible–NIR region 400–1100 nm (exposed and non-exposed sides of the fruit averaged and MSC pre-processed) and expressed as 'Absorbance' for (a) 'Aroma' and (b) 'Holsteiner Cox' in two measurement dates. Axes: X = wavelength (400–1100 nm), and Y = absorbance.

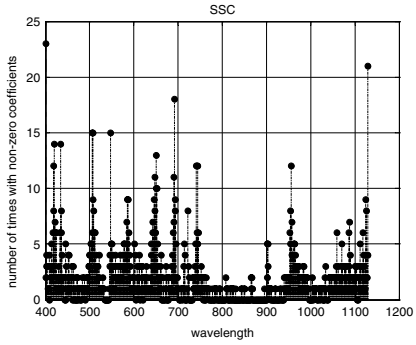


Figure 3. The frequency of having non-zero regression coefficients in 25 iterations of EN for the original SSC data set for 'Aroma' apples.

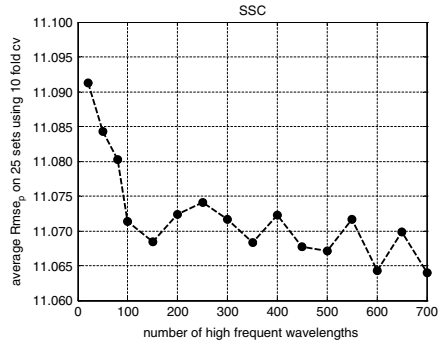


Figure 4. The RMSEP candidate number for SSC.

All the described steps were also performed with the MSC pre-processed data. The number of selected bands for SSC and acidity in this case were 450 and 250 respectively.

Results for the calibration and validation models

The resulting numbers of fruit samples on each of the previously explained sampling arrangements were: Sets A and C of 'Aroma' had 147 and 49 (196 in total) samples for SSC and 141 and 48 (189 in total) samples for acidity. 'Holsteiner Cox' SSC training and test sets A and C had 165 and 54 (219 in total) respectively and the sets for acidity had 152 and 51 (203 in total) samples.

The smallest RMSEPs from each combination of the three arrangements and two data sets (original/MSC) are presented in Fig. 6. The selected features on Set A (smooth arrangement) using the EN regression and 26 wavelengths obtained the best

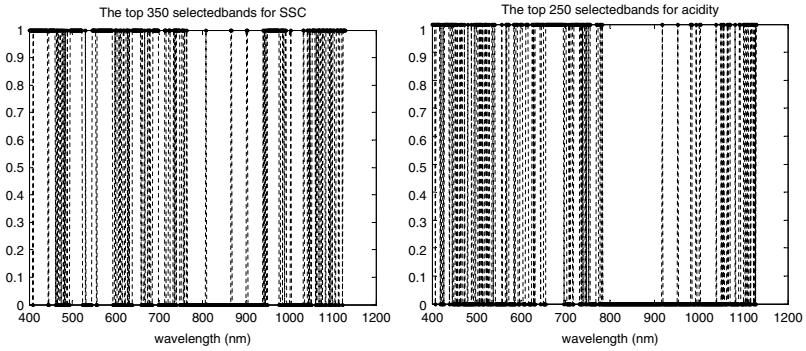


Figure 5. Stem plots for the selected bands for SSC and acidity for the original data set.

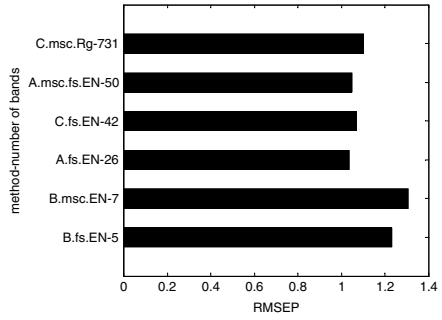


Figure 6. Overall comparison of the soluble solids content (SSC) prediction errors based on RMSEPs of the three training/test arrangements (A, B and C respectively) and the two original (without) and MSC pre-processed data. MSC, multiplicative scatter correction; Rg, ridge regression; FS, feature selection; EN, elastic net. The numbers on the ordinate indicate the number of wavelengths used in the models.

results for SSC and acidity prediction. Therefore complete results of sets B and C are not presented.

Figure 7 illustrates a comparison of the error of the prediction models obtained on sets B (Fig. 7a) and C (Fig. 7b) respectively. The figure further demonstrates the importance of the strategy used for forming the training and test sets. The minimum RMSEPs for set B were higher than the worst results obtained using set A using PLS. Set C also produced better models than Set B, but the best results were not as good as those for Set A.

Sets B and C had higher prediction errors. In particular the random sets (C) often caused over fitting during the modelling process, resulting often in poorer models. As an illustration, Fig. 8 shows the spread of RMSEP from the three sample arrangements used for building the prediction models of SSC of 'Aroma' apples. Summary statistics for the four sets and their average formed during smooth arrangement (Set A) are shown in Table 1.

Table 2 presents the average prediction statistics for the set 'A' of the 'Aroma' cultivar, which obtained the best results for both SSC and acidity models. In general, the calibration and prediction correlation using PLS were inferior to RR and EN in all cases (Table 2 and Table 3). The error and bias remained low for set A.

In a similar manner, Table 3 shows the ridge and EN regression results for the selected bands of the original and MSC pre-processed data. For the SSC data, in all cases except ridge regression on the original data, the performance slightly improved using the reduced number of wavelengths. In the case of acidity, the effect was the same.

Soluble solids content

Table 2 shows that elastic net and ridge regression improved the prediction of SSC and also resulted in lower errors and bias as compared to PLS. Table 3 demonstrates that even though all the models performance in terms of R^2, errors and bias were not importantly improved after band selection, fewer bands on the visible and NIR region were suited for SSC prediction for set A. Fewer bands simplify the measurement systems and make them more cost effective. Our 'Aroma' SSC calibration models performed better ($R^2 = 0.44$; SEC = 0.88°Brix; range: 8.0–15.5°Brix; SD = 1.39) than previous studies done by Zude *et al.*,[21] which reported $R^2 = 0.04$ and higher SEC = 1.82 (for stored 'Golden Delicious' apples). They used higher number of samples in storage ($n = 250$; SD = not reported) and spectral bands between 400 and 1000 nm. On the other hand, Dai *et al.*[24] obtained SSC prediction models with higher R^2 using smaller sample numbers ($N = 58$; $R^2 = 0.76$; SEC = 0.22; SEP = 0.83), similar band range 400–1100 nm for a data set with SSC values fairly normally distributed around the mean ($range_{cal} = 8.6–16.7$; $range_{val} = 8.6–15.5$ $SD_{cal} = 1.69$; $SD_{val} = 1.62$). The high difference between SEC and SEP in this latter study indicates that the training and test samples were not very similar. Another model from Hernández *et al.*[25] also had high prediction results ($R^2 = 0.98$; SEC = 0.45°Brix; SEP = 1.69°Brix) except bias = 1.62 was quite high. They used samples with higher variation than ours ($CV_{SSC} = 0.28$). It was not clear, however, how the samples were collected in these latter studies. Peirs *et al.*[8] calculated a SSC model using 244 apple samples for calibration and 244 samples validation from seven different apple cultivars where $R^2_{cal} = 0.91$ to 0.92, SEC = 0.49 to 0.76 but using spectral regions between 380 and 2000 nm. It is possible that better results for SSC might be obtained by extending the spectral data to other areas of the NIR spectra to the ones we studied.

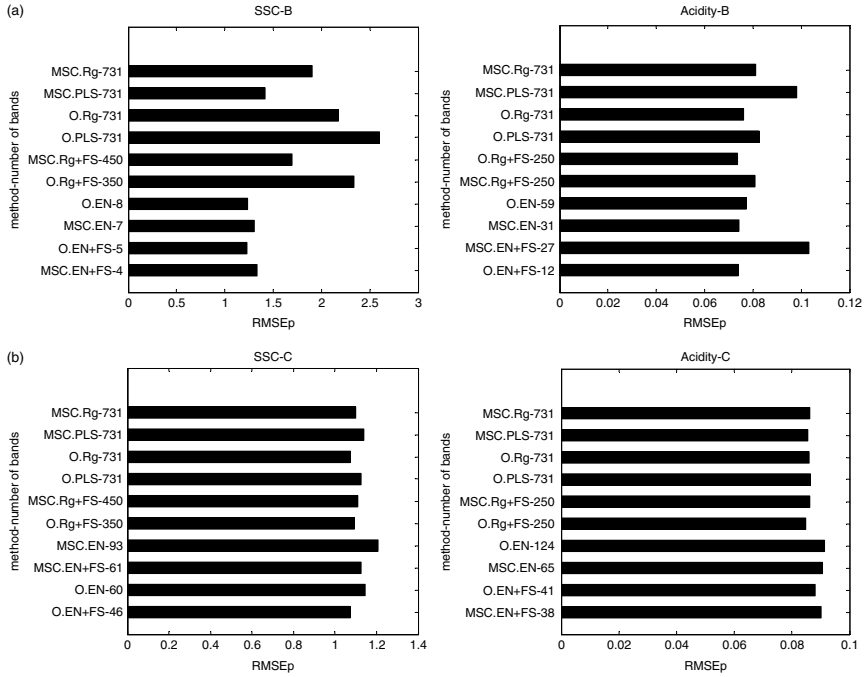


Figure 7. Comparison of the RMSEPs of the different regression methods used on the data division for modelling sub-samples (a) Set B and (b) set C for 'Aroma' cultivar. The numbers on the ordinate indicate the number of wavelengths used in the models and the letters indicate, in each case, the approach used to analyse the data sets. MSC, multiplicative scatter correction; O, original data without MSC; Rg, ridge regression; FS, feature selection; EN, elastic net.

Table 1. Statistics for the training/test sub-sample sets for the sample arrangement A

Characteristic	Statistic	Set 1	Set 2	Set 3	Set 4	Average
Soluble solids content (°Brix)	Number	147/49	147/49	147/49	147/49	147/49
	Range	7.9–14.4/8.3–15.9	7.9–15.9/8.2–14.3	7.9–15.9/8.2–14.4	8.2–15.9/7.9–14.1	8.0–15.5/8.1–14.7
	Mean	11.05/11.08	11.06/11.05	11.06/11.04	11.06/11.05	11.1/11.1
	SD	1.37/1.47	1.41/1.37	1.40/1.38	1.40/1.39	1.39/1.40
Acidity (g 100 mL^{-1})	Number	141/48	141/48	141/48	141/48	141/48
	Range	0.5–1.1/0.45–0.87	0.47–1/0.45–0.88	0.45–0.91/0.5–0.1	0.45–1/0.51–0.9	0.47–0.98/0.48–0.91
	Mean	0.67/0.67	0.67/0.67	0.67/0.67	0.67/0.67	0.67/0.67
	SD	0.09/0.01	0.09/0.1	0.09/0.1	0.1/0.09	0.09/0.1

Acidity

Our acidity models performed lower those reported by Peirs *et al.*[8] They used random samples from a combination of seven different apple cultivars to construct calibration ($N = 244$) and validation ($N = 244$) models in the region between 380 and 2000 nm. Their results were $R^2_{cal} = 0.88$, $R^2_{val} = 0.86$ (SEC = 1.64; SEP = 1.73). Abu-Khalaf and Bennedsen[26] reported also better results using in total 200 samples of two apple cultivars ('Golden Delicious' and 'Jonagold') to calculate calibration ($N = 130$; $r^2 = 0.84$; SEC = 0.07) and validation models ($N = 70$; SEP = 0.07) on the spectral region 400–1100 nm. Other studies have reported improved prediction results for acidity on citrus fruit using the spectral region between 500–1100 nm ($r^2 = 0.65$; RMSEP = 0.15) and up to the 2500 nm acidity predictions have reached $r^2 = 0.86$; RMSEP = 0.17.[27]

'Holsteiner Cox' models had poor prediction, which was somehow expected because the sample variability from this cultivar was very low (CV of 5–6%). However, this also means that the spectra may be did not capture SSC or acidity levels accurately for this cultivar.

Table 2. Averaged calibration and prediction results (Set A) for 'Aroma' apple cultivar using all 731 spectral bands

		Raw data sets			MSC transformed data sets		
Characteristic	Statistic	PLS	RR	EN	PLS	RR	EN
Soluble solids content (°Brix)	N_{cal}	147	147	147	147	147	147
	N_{val}	49	49	49	49	49	49
	RMSECV	0.91	0.78	0.90	0.91	0.87	0.85
	RMSEP	1.09	1.04	1.04	1.11	1.08	1.05
	R^2_{cal}	0.55	0.68	0.56	0.60	0.60	0.61
	R^2_{val}	0.33	0.43	0.43	0.35	0.39	0.43
	SEC	0.93	0.79	0.92	0.92	0.88	0.86
	SEP	1.07	1.04	1.04	1.11	1.08	0.99
	$Bias_{cal}$	0.00	0.00	0.00	0.00	0.00	0.00
	$Bias_{val}$	0.01	0.02	0.01	−0.03	0.01	0.02
	NNC	731	731	45	731	731	134
Acidity (g 100 mL^{-1})	N_{cal}	141	141	141	141	141	141
	N_{val}	48	48	48	48	48	48
	RMSECV	0.08	0.08	0.07	0.08	0.08	0.08
	RMSEP	0.08	0.08	0.08	0.09	0.08	0.08
	R^2_{cal}	0.29	0.30	0.36	0.29	0.27	0.33
	R^2_{val}	0.22	0.22	0.22	0.15	0.18	0.18
	SEC	0.08	0.08	0.08	0.08	0.08	0.08
	SEP	0.09	0.09	0.08	0.09	0.09	0.09
	$Bias_{cal}$	0.00	0.00	0.00	0.00	0.00	0.00
	$Bias_{val}$	−0.00	−0.00	−0.00	−0.00	−0.00	−0.00
	NNC	731	731	81	731	731	26

MSC, multiplicative scatter correction.
PLS, partial least squares method.
RR, ridge regression.
EN, Elastic Net.
N_{cal}, number of samples in the calibration set.
N_{val}, number of samples in the prediction set.
RMSECV, root mean square error of cross validation.
RMSEP, root mean square error of prediction.
R^2_{cal}, coefficient of determination (calibration).
R^2_{val}, coefficient of determination (validation).
SEC, standard error of calibration.
SEP, standard error of prediction.
$Bias_{cal}$, bias calibration.
$Bias_{val}$, bias validation.
NNC, number of non-zero coefficients.

The differences on model robustness observed between the three approaches used for forming training and test sets, indicated that training and test sample arrangement do affect the overall model performance, especially when the number of samples are limited and smaller than the number of wavelengths. The tendency of producing stable prediction after applying the smooth fractionator to form training and test sets, stressed the importance of maintaining the original sample variability throughout the entire model construction process in order to achieve model robustness and high precision (higher coefficients of determination, low errors between calibration and validation sets and low bias). This conclusion is supported by the differences observed between our training and test sets performance, which suggests that the variability of the training set were, by chance, excluded from the test set during the formation of training and test sets. Consequently, using a different sampling period ($p = 2$) to form training and test sets, so that the proportion becomes 50–50 instead of the commonly used 75–25, might be a better approach to use in order to preserve as much as possible the original variability of the whole data set, when working with smaller samples. The fractionator technique used to sample fruit from the trees has already been shown to be an effective way to obtain small samples (<100) representing the variability of fruit size and internal quality at the orchard scale, as shown by Wulfsohn *et al.*[28] and Martinez V *et al.* (unpublished). These previous studies and the findings of this study suggest that is feasible to develop robust visible–NIR prediction models using relatively small samples. The type of cross validation used might have some additive effect on the model performance as well, because it is a method also based on repetitive selection of samples from the calibration set.[29] However, given the low results for the model errors and bias, we consider the robustness of our models to be adequate for this type of data set.

In practice, the results of this study imply that using local fruit samples for developing spectral calibration and prediction sets is feasible, as long as the sample variability is taken into account on the formation of training and prediction sets. The late season samples, however, need to be modelled differently. Using representative fruit samples with higher internal quality variability (e.g. CV > 15%) might be a better approach to use, since a much

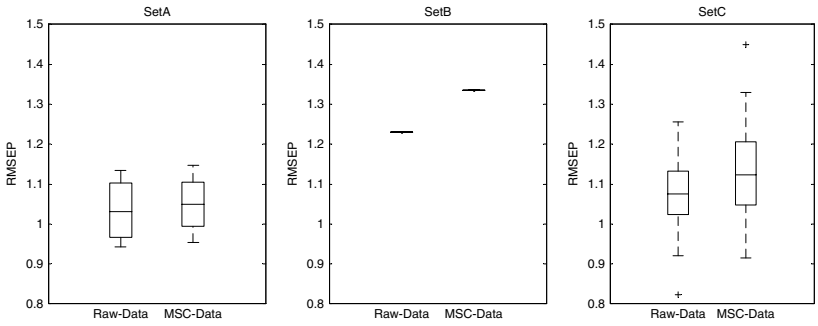


Figure 8. Box plots for the root mean square of prediction (RMSEP) for the raw and MSC treated data of the sample arrangements A, B and C.

Table 3. Averaged calibration and prediction results (set A) for Aroma apple cultivar using selected bands

		Raw data sets			MSC transformed data sets		
Characteristic	Statistic	PLS	RR	EN	PLS	RR	EN
Soluble solids content (°Brix)	N_{cal}	147	147	147	147	147	147
	N_{val}	49	49	49	49	49	49
	RMSECV	0.87	0.77	0.88	0.92	0.88	0.87
	RMSEP	1.14	1.07	1.03	1.13	1.07	1.05
	R^2_{cal}	0.60	0.69	0.60	0.55	0.60	0.61
	R^2_{val}	0.33	0.40	0.44	0.33	0.40	0.43
	SEC	0.88	0.78	0.88	0.93	0.89	0.87
	SEP	1.14	1.07	1.03	1.13	1.08	1.05
	$Bias_{cal}$	0.00	0.00	0.00	0.00	0.00	0.00
	$Bias_{val}$	0.04	0.03	0.03	0.00	0.02	0.02
	NNC	350	350	26	450	450	50
Acidity (g 100 mL^{-1})	N_{cal}	141	141	141	141	141	141
	N_{val}	48	48	48	48	48	48
	RMSECV	0.08	0.08	0.08	0.08	0.08	0.08
	RMSEP	0.08	0.08	0.09	0.08	0.08	0.08
	R^2_{cal}	0.32	0.31	0.27	0.32	0.31	0.30
	R^2_{val}	0.22	0.24	0.19	0.19	0.20	0.18
	SEC	0.08	0.08	0.08	0.08	0.08	0.08
	SEP	0.09	0.08	0.09	0.08	0.08	0.09
	$Bias_{cal}$	0.00	0.00	0.00	0.00	0.00	0.00
	$Bias_{val}$	−0.00	−0.00	−0.00	−0.00	−0.00	−0.00
	NNC	250	250	37	250	250	19

Abbreviations as in the footnote to Table 2.

higher number of samples of the fruit populations with the same variability as the samples we presented will not necessarily improve the prediction results greatly. Such high variability samples may be obtained from commercial orchards where the internal variability of fruits is expected to be higher. We are testing this hypothesis in different batches in our further research.

CONCLUSION

Three sub-sampling techniques for the formation of training and test sets for spectral prediction of SSC and acidity of an early apple cultivar were compared. The smooth fractionator approach was clearly superior to random sampling and to by-date separation, resulting in models with consistently low errors and low bias, mainly because the method provides a fair representation of the response values in both the training and test sets. In addition, three different methods to reduce model complexity, PLS, RR and EN were compared. Using elastic net and fewer bands were the best approaches to reduce model complexity ($R^2 = 0.44$; SEP = 1.03°Brix; bias = 0.03; range: 8.1–14.7°Brix). Therefore our results confirmed that the fractionator sampling provide data sets suitable for SSC prediction with visible–NIR spectroscopy. To our knowledge, this is the first proposal of a modelling protocol for a sub-sample of training and test sets, which takes into account the variability of the original sample in the context of predicting fruit quality by using a non-destructive method.

ACKNOWLEDGEMENTS

We express our gratitude to the funding body for this research: the Danish Ministry of Food, Agriculture and Fisheries (Food research program 2008: ''Ydun Juice: Potential for specialty juices from Danish local apple cultivars'') and to our project partner NordGen. Also we would like to thank Jesper Dauw for his valuable input during the pre-processing of raw data. The assistance of the Pometum staff is highly appreciated.

REFERENCES

1 Cozzolino D, Cynkar WU, Shah N and Smith P, Multivariate data analysis applied to spectroscopy: potential application to juice and fruit quality. *Food Res Int* **44**:1888–1896 (2011).
2 McCaffrey D, Harvesting the sun. A profile of world horticulture. *Script Hortuc* **14** (2012).
3 Nicolai B, Beullens K, Bobelyn E, Peirs A, Saeys W, Theron K, *et al.*, Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biol Technol* **13**:99–118 (2007).
4 Mehinagic E, Royer G, Bertrand D, Symoneaux R, Laurens F and Jourjon, Relationship between sensory analysis, penetrometry and visible–NIR spectroscopy of apples belonging to different cultivars. *Food Qual Prefer* **14**:473–484 (2003).
5 Mendoza F, Lu R and Cen H, Comparison and fusion of four non-destructive sensors for predicting firmness and soluble solids content. *Postharvest Biol Technol* **73**: 89–98 (2012).
6 Bobelyn E, Serban A, Nicu M, Lammertyn J, Nicolai B and Wouter S, Postharvest quality of apple predicted by NIR-spectroscopy: study of the effect of biological variability on spectra and model performance. *Postharvest Biol Technol* **55**:122–143 (2010).
7 McGlone VA and Kawano S, Firmness, dry-matter and soluble solids assessment of postharvest kiwifruit by NIR spectroscopy. *Postharvest Biol Technol* **13**:131–141 (1998).
8 Peirs A, Lammertyn J, Ooms K and Nicolai BM, Prediction of the optimal picking date of different apple cultivars by means of VIS/NIR spectroscopy. *Postharvest Biol Technol* **21**:189–199 (2000).
9 Wulfsohn D, Maletti G and Toldam-Andersen TB, Unbiased estimator of the total number of flowers on a tree. *Acta Hortic* **707**:245–251 (2006).
10 Næs T, Isaksson T, Fearn T and Davies T, *A User-Friendly Guide to Multivariate Calibration and Classification*. NIR Publications, Chichester, pp. 114–115 (2002).
11 Gundersen HJG, The smooth fractionator. *J Microsc* **207**:191–210 (2002)
12 Wulfsohn D, Sampling techniques for plants and soil. *Landbauforschung Volkenrode* **340**:3–30 (2010).
13 Hastie T, Tibshirani R and Friedman J, *The Elements of Statistical Learning*, 2nd edition. Springer, Canada, pp. 61–68 (2009).
14 Zou H and Hastie T, Regularization and variable selection via Elastic Net. *J R Statistics Soc* **67**:301–320 (2005).
15 Hoyle D C and Rattray M, Statistical mechanics of learning multiple orthogonal signals: Asympotic theory and fluctuation effects. *Phys Rev E (Statistical, Nonlinear and Soft Matter Physics)*, **75**:016101 (2007).
16 Sharifzadeh S, Clemmensen HL, Borggaard C, Støier S and Ersbøll KB, Supervised Feature Selection for Linear and Non-Linear Regression of L*a*b* Color from Multispectral Images of Meat, Accepted at Eng App Artif Intel, Manuscript Number: EAAI-12-1827 (2013).
17 Korsgaard M and Toldam-Andersen TB, *The Apple Key*. Available: http://www.nordgen.org/nak/index.php?view=show&id=5003 [7 August 2012].
18 Qing Z, Bi J and Zude M, Wavelength selection for predicting physicochemical properties of apple fruit based on near-infrared spectroscopy. *J Food Qual* **30**:511–526 (2007).
19 Xiaobo Z, Yanxiao L and Jiewen Z, Using genetic algorithm interval partial least squares selection of optimal near infrared wavelength regions for determination of the soluble solids content of 'Fuji' apple. *J Near Infrared Spectrosc* **15**:153–159 (2007).
20 Rizzolo A, Vanolli M, Spinelli L and Torricelli A, Sensory characteristics, quality and optical properties measured by time reflectance spectroscopy in stored apples. *Postharvest Biol Technol* **58**:1–12 (2010).
21 Zude M, Herold B, Roger JM, Bellon-Maurel V and Landahl S, Non-destructive tests on the prediction of apple fruit flesh firmness and soluble solids content on tree and in shelf life. *J Food Eng* **77**:254–260 (2006).
22 Sun D, *Contemporary Food Engineering*. CRC Press, New York, chapter 3 (2009).
23 Chauchard F, Roger JM and Bellon-Maurel V, Correction of the temperature effect on near infrared calibration–application to soluble solid content prediction. *J Near Infrared Spectrosc* **12**:199–205 (2004).
24 Dai F, Hong T, Zhang K and Chen H, Comparison of modelling methods in rapid estimation of sugar content of apple based on near infrared spectrum. An ASABE meeting presentation, paper number: 1008620. ASABE, St. Joseph, Mich. (2010).
25 Hernández N, Lurol S, Roger J and Bellon-Maurel V, Robustness of models based on NIR spectra for sugar content prediction in apples. *J Near Infrared Spectrosc* **11**:97–107 (2003).
26 Abu-Khalaf N and Benenedsen B, Near infrared (NIR) technology and multivariate data analysis for sensing taste attributes of apples. *Int Agrophys* **18**:203–211 (2004)
27 Magwaza LS, Opara UL, Niewoudt H, Cronje PJR, Saeys W and Nicolaï B, NIR spectroscopy applications for internal and external quality analysis of citrus fruit – A review. *Food Bioproc Technol* **5**:425–444 (2012).
28 Wulfsohn D, Aravena F, Potin C, Zamora I and Garcia-Fiñana M, Multilevel systematic sampling to estimate total fruit number for yield forecasts. *Precis Agric 13*, **2**:256–275 (2012).
29 Gemperline P, *Practical Guide to Chemometrics*, 2nd edition. Taylor and Francis, Boca Raton, pp. 115–116 (2006).

Appendix F

Statistical quality assessment of pre-fried carrots using multispectral imaging

Authors: Sara Sharifzadeh[1], Line H. Clemmensen[1], Hanne Løje[2], Bjarne K. Ersbøll[1]

1. Department of Applied Mathematics and Computer Science, Technical University of Denmark.
2. National Food Institute, Technical University of Denmark.

Published in *SCIA 2013 LNCS 7944 proceedings, pp. 620-629, 2013, Springer-Verlag, 2013.*

Statistical Quality Assessment of Pre-Fried Carrots using Multispectral Imaging

Sara Sharifzadeh[1*], Line H. Clemmensen[1], Hanne Løje[2], Bjarne K. Ersbøll[1]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark
{sarash,lkhc,bker}@dtu.dk
[2]National Food Institute, Technical University of Denmark
halo@food.dtu.dk

Abstract.

Multispectral imaging is increasingly being used for quality assessment of food items due to its non-invasive benefits. In this paper, we investigate the use of multispectral images of pre-fried carrots, to detect changes over a period of 14 days. The idea is to distinguish changes in quality from spectral images of visible and NIR bands. High dimensional feature vectors were formed from all possible ratios of spectral bands in 9 different percentiles per piece of carrot. We propose to use a multiple hypothesis testing technique based on the Benjamini-Hachberg (BH) method to distinguish possible significant changes in features during the inspection days. Discrimination by the SVM classifier supported these results. Additionally, 2-sided t-tests on the predictions of the elastic-net regressions were carried out to compare our results with previous studies on fried carrots. The experimental results showed that the most significant changes occured in day 2 and day 14.

Keywords: Multispectral imaging, Multiple hypothesis testing, Segmentation, Food quality assessment, SVM classification, Elastic-net regression

1 Introduction

Spectral imaging has recently gained use in on-line quality monitoring of food items. This method has important privileges over the traditional assessment methods, based on skillful human experts. The imaging-based methods are automatic, fast and contact-less. Since, there is a trend toward the use of these methods for quality control of food products such as meat, dairies and vegetables; it is of importance to test the capabilities as well as the reproducibility of such.

In this paper, the multispectral images of pre-processed carrots were used to detect the effect of storage on their color and NIR characteristics. The carrots were pre-fried without oil and then frozen for about two months. Then, they were moved to the refrigerator for experiments during a period of 14 days. Generally the surface color and texture are important parameters; indicating the quality of food. Multispectral images provide this information in visible bands and also more information about the subsur-

face and chemical characteristics in NIR bands. Using the multispectral images in visible and NIR bands, we tracked the quality of carrots during the storage days. The aim was to find out in which days, significant changes occurred.

In this study, the preparation of carrots was performed in two steps. First, the vegetables were stir-fried (without oil) [1, 2]. Research findings have shown that, stir-frying produces high quality vegetables [1]. After stir-frying, the products were frozen. The concept of partial thawing during distribution which can improve shelf life was used [3, 4].

Previous studies showed that, multispectral images can be used to assess the color change over time in pre-fried vegetables [5]. The analysis of multispectral images of pre-fried carrots and celeriac (fried in oil) was carried out for change detection with in a period of 14 days in [6, 7]. In [6], the segmented vegetable pieces were used to form the high dimensional feature vectors (3249 variables). For each carrot piece, 9 different percentiles in all possible ratios of bands were calculated. Then, the elastic-net regression analysis was carried out to predict the days of analysis from these high dimensional features. Finally, a 2-sided t-test on the estimated days was applied. A significant change was detected in carrots from day 2 to 4 at a 5% level of significance. In [7], instead of the ratios of bands, the feature vectors were calculated using the percentiles of pixel intensity values within each vegetable piece. These features were used directly for unpaired t-tests to detect trends of change in reflection, as a function of days kept in the refrigerator. Again for carrot, the significant change was detected from day 2 to 4 at 5% level. The results were also compared with the sensory panel tests.

In this paper, we propose to apply a multiple hypothesis testing technique to assess the high dimensional features obtained from the ratios of spectral bands and their corresponding percentiles. The high dimensional features based on band ratios are preferred, since they are more robust toward the undesired effects such as shadows. Multiple hypothesis techniques are mostly used for genomic data [8, 9]. They involve the significance assessment of the individual features. This assessment was performed, without the use of multivariate predictive models like in [6]. Since the dimensionality of the extracted features was quite high (3078), a conventional t-test at a significance level e.g. α=0.05 may find about 154 significant features just by chance even if, the null hypothesis of no change is true for all the features [9]. In our study, the False Discovery Rate (FDR) introduced in [11] and the expected number of significant features was used to detect the significant days of change. In addition, the Support Vector Machine (SVM) classification was employed. Although the classification results support the multiple hypotheses testing, it is difficult to use them alone, as a demonstration for significance of changes over the days. In addition, the method used in [6] was applied to our data set, and the results were compared with the findings from the multiple hypothesis tests. Finally, we found the wavelengths mostly represented the significant features over the inspection days.

The rest of this paper is organized as follows; section 2 describes the data preparation for the experiments and the feature extraction step. In section 3, we explain the data analysis and section 4 presents the results. Finally, there is a conclusion for this paper in section 5.

2 Data Preparation and Feature extraction

2.1 Data Preparation

The carrots used in this study, were prepared for the experiments a few days after harvest. First, they were cut into cube shapes of size one cm^3 approximately. Then, they were wok-fried for 4 minutes at 250° C in the continuous stir [1]. After cooling down, they were packed in 500g weight in plastic bags and frozen at -30° C for 50-60 days. Finally, they were moved to the refrigerator (+5° C). On days 2, 5, 8, 11 and 14 one bag was taken out from the refrigerator and the imaging experiments were conducted. The VideometeLab[1] was used for multispectral imaging like in [6, 7]. The multispectral images were obtained in 19 different wavelengths with the image resolution of 1280×960 pixels. The spectral images of a sample petri dish of carrots are shown in Fig.1 and pseudo-RGB images of samples in the five inspection days are shown in Fig 2.

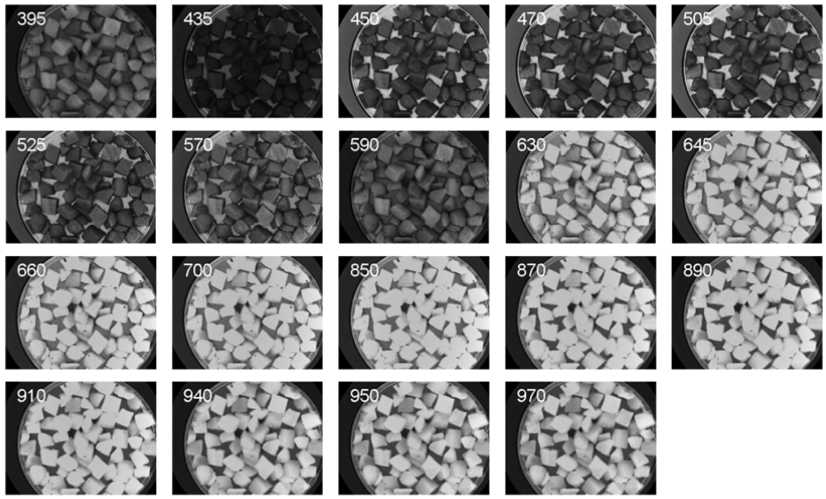


Fig. 1. Spectral images of carrot in 19 wavelengths, shown in nanometer range.

Fig. 2. Pseudo-RGB images of carrot samples. From left to right: day 2, 5, 8, 11, 14

[1] www.videometer.com

2.2 Feature extraction

In order to form feature vectors, the first step is to segment the carrot pieces from the background and also from each other. Some of the procedures are almost similar to those in [6]. A brief description of different steps is presented in the following.

The background, which is everything except carrot, was removed in three different steps. First, a simple thresholding was performed manually, considering the histograms of the images. But, some parts of the background that had intensity variations close to the carrots, still remained. Therefore, two populations were considered (carrot and remaining background) and labeled manually. Then Canonical Discriminant Analysis (CDA) [10] was employed to improve the discrimination level between the two populations to define better thresholds [12]. Finally, a fine morphological operation (closing and erosion) was used to clean the undesired small remaining areas. For segmentation of carrot pieces, similar to [13], the Sobel edge detector was used which is based on the gradient function. Then, some morphological operations followed by a Watershed transformation were applied on the background-removed image. Using both the gradient and Watershed transform, the carrot pieces were segmented.

Each carrot piece was considered as a single observation. Since there were lumps of pieces in the petri dishes instead of individual well separated cubes, the segmentation was not perfect. In some cases two carrots were segmented as one piece or one piece was segmented into two different pieces. However, all segmented areas include carrots bodies. Fig.3 illustrates different steps of the carrot segmentation. A feature vector was formed for each detected carrot piece. First, for each of the 19 bands, all its 18 possible ratios to the other bands were calculated (totally 342 ratio matrices). Then, in each ratio matrix, 9 percentiles (1, 5, 10, 25, 50, 75, 90, 95, 99) were calculated for each piece. Therefore, by concatenating of the 9 percentile vectors of the 342 ratios, a 3078 length feature vector per piece of carrot was obtained.

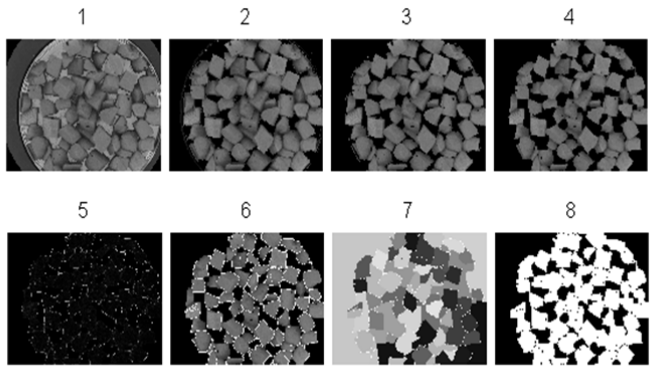


Fig. 3. Segmentation steps: 1-original image 2-two manual thresholding steps 3-thresholding using CDA 4-morphological operation 5-detected edges by Sobel 6-detected edges by Watershed 7-Watershed label matrix 8-segmented pieces

3 Data Analysis

As mentioned before, the aim of the analysis is to detect and track changes in the color and NIR characteristics of the carrots over the days of inspection. Totally, 3277 observations (carrot pieces) were segmented from the images of the 5 days of experiments. They were divided into two sets (*set1* and *set2*) randomly and all analyses were performed twice, using one of the sets as training and the other as test set and vice versa.

3.1 Elastic-net Regression and 2-sided t-test

Looking to the problem from a statistical point of view, a statistical test can be used for finding the significant changes. For this aim, like in [6], an elastic-net regression was performed on the ratio features to predict the number of days. 10-fold cross validation was used to generalize the prediction model. In the next step, a 2-sided t-test was applied on the predicted labels for all pairs of days to test the null hypothesis H_0 that two groups come from the same population at the 5% level of significance. This means that instead of considering the high dimensional features directly, they were first used for predicting the days and the one–dimensional prediction vector was used for the statistical test.

3.2 SVM Classification

The main problem at hand can be considered as discrimination of the data between different labels (days). If we look at the problem from this point of view, a classification approach can be used as a means of discrimination. For this aim, the powerful, support vector machine classifier was employed using LIBSVM [13]. A linear kernel with 5-fold cross validation was used for training the classifier. We expect that in days where considerable changes occurred, most of the samples be truly classified. On the other hand, for days with more similarity, misclassified samples should increase. Therefore, we will look at the confusion matrix for classification between days.

3.3 Multiple Hypothesis Testing

Another way of performing a statistical test is to perform multiple hypothesis tests for individual features of pairs of days to find the significant changes. However, not all the detected significant features are truly significant. For example, if the conventional statistical t-test with a priori significance level of α=5% be used, just by chance, about 154 significantly changed features will be found out of the total 3078 features, while the null hypothesis may be true for some of them. Table 1 shows the theoretical outcomes from the M hypothesis tests [10]. Where, M is the number of features.

To address this problem, the number of falsely significant features (V in Table 1), for which the null hypothesis of no change (H_0) is true, should be found. One simple

Table 1. Possible outcomes from M hypothesis tests.

	Called not Significant	*Called Significant*	*Total*
H0 True	U	V	M_0
H0 False	T	S	M_1
Total	M-R	R	M

solution is the *Bonferroni* method. In order to reduce the number of false positive features (V), it rejects H_0 if the p-value of a feature satisfies $p < \alpha/M$. It is a useful method in cases that M is small, as it is based on the assumption that the covariates are independent. However, in our case M is quit high (M=3078) and high correlation exists between the covariates. Therefore, the *Benjamin-Hachberg* (BH) method is used instead. They introduced the False Discovery Rate (FDR) as follows:

$$FDR = E(\frac{V}{R}) \tag{1}$$

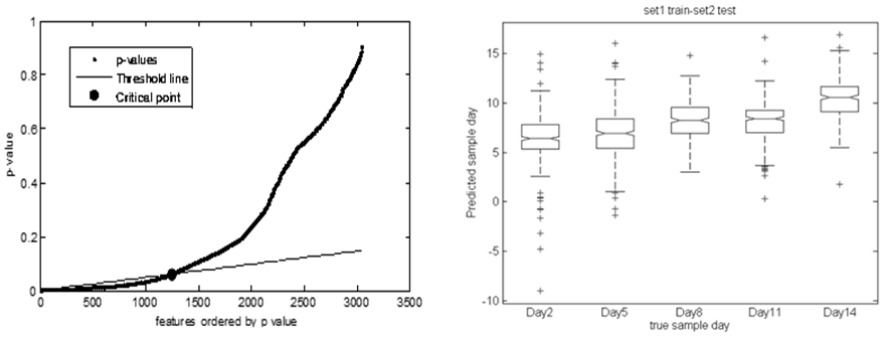
It is the expected proportion of the false positive features V among the R features that are called significant. In this method, the FDR rate is bounded by a user defined level α. It is calculated based on the p-values obtained from an asymptotic approximation of the test statistic like a Gaussian or a permutation distribution.

In this paper, a *plug- in* version of this method is followed [10]. The FDR rate is bounded at α=0.15 and the permutation distribution is used. Then, the number of truly significant features $\widehat{E(S)}$ is estimated and used for making decision about the days that significant changes occurred in the vegetable data. The procedures are as follows:

1. The t statistics are calculated for all the features of both days: t_j , $\mathrm{j} = 1, \dots, \mathrm{M}$
2. For K=1000 times, the sample's labels (days) are permuted and the t statistic for the features at each permutation round is calculated. t_j^k, $\mathrm{j} = 1, \dots, \mathrm{M}$, $\mathrm{k} = 1, \dots, \mathrm{K}$
3. For a cut-point C, the $R = \sum_{j=1}^{M} I(|t_j| > C)$, as well as the $\widehat{E(V)} = \frac{1}{K}\sum_{j=1}^{M}\sum_{k=1}^{K} I(|t_j^k| > C)$ are calculated. (the I function shows the number of times the inequality is satisfied)
4. The $\widehat{FDR} = \widehat{E(V)}/R$ and the number of truly significant features $\widehat{E(S)} = R - \widehat{E(V)}$ are computed.

In this study, the cut-point C is chosen equal to the t statistic of the critical point of the BH method. The critical value controls the FDR to be around α=0.15. It is calculated in the following steps:

1. The corresponding pooled p-value of the t statistics for each feature is computed: $p_j = \frac{1}{MK}\sum_{j'=1}^{M}\sum_{k=1}^{K} I(|t_{j'}^k| > |t_j|)$
2. The p-values for the features are ranked in ascending order $p_{(1)} \le p_{(2)} \le \cdots \le p_{(M)}$
3. The BH critical point is $L = max\{j: p_{(j)} < \alpha \times j/M\}$ and so the cut point is C=$|t_L|$.



(a) (b)

Fig. 4. (a) A plot of the ordered p-values $p_{(j)}$, the threshold line ($\alpha \times j/M$) as well as the critical point of the BH method. (b) The box-plots of the elastic-net prediction, set1 training set2 test.

Fig.4 (a) shows a sample plot of the ordered p-values, the threshold line ($\alpha \times j/M$) as well as the critical point. Using this method, the number of significantly changed features is obtained for all pairs of days.

4 Results and Discussion

In the first experiment, like in [6], an elastic-net prediction was performed using the spectral data of set1 to train the prediction model. Then, set2 was used for test. Fig. 4(b) shows the boxplot of the result. The same experiment was repeated, when the training and test sets were swapped. The prediction error was high (similar to [6]) and the test MSEs were 13.55 and 13.66 for the two experiments respectively. Although the prediction accuracy was poor, a linear trend could be seen for the predictions over the 5 days. A pairwise 2-sided t-test (at a 5% level of significance) was performed for all pairs of days using the prediction output. The results of both sets showed significant change almost between all pairs of days.

In the next experiment, SVM classification was performed. Table 2 shows the confusion matrix of the SVM test results, where set2 was used for training the model and set1 was used for test. It shows the percentage of each day's samples that were classified into one of the 5 days of inspection. Therefore, all rows sum to 100 percent. As can be seen, at each day, more than half of the observations were classified to the same class (inspection day) correctly, and less observations were misclassified. This implies the capability of the multispectral imaging to distinguish subtle changes in carrots over days that are even difficult to be observed by eyes, as we have seen in Fig. 2.

On the other hand, the number of misclassified samples was also considerable in many cases which illustrate the similarity of the samples in different days. Day 14 gained the minimum misclassification and 78.66% of the samples were just truly classified in that day. For the second experiment, where the training and test sets were

swapped, the same holds true for day 14. Day two gained the second highest classification performance. However, in the second experiment its percentage was around 60.29. Therefore, more information about the significance of the changes seemed necessary to confirm a change in quality of carrots between days.

At this step, we used the results from the multiple hypothesis tests explained in the previous section. Tables 3, shows the number of significantly changed features in all pairs of days for set1. These results were obtained using the BH method for the FDR rate bounded at 0.15. The Table is symmetric to the diagonal. The pairwise analysis between day 14 and the all other days showed the highest number of significantly changed features. This is compatible with the previous results from the SVM classification. The next highest number of significant features was detected from the pairwise analysis of day 2 with the other days. This was seen in one of the SVM test results as well. The same analysis was performed on set2 and the results were similar to set 1. Therefore, we can conclude from these analyses that, the most important change in carrot samples occurred after 2 weeks. While with less significance level, they also changed after 2 days being kept in the refrigerator.

Comparison of the results obtained from the SVM and BH methods with those from the t-test on the elastic-net predictions, showed differences in the significant days of change. However, we believe that the SVM and BH methods results are more reliable. First, since they were obtained by the direct use of the 3078 features of spectral image than the prediction results. Second, the elastic-net regression prediction error was high.

Table 2. The percentage of each day data assigned to the 5 classes by SVM

	Class 2	*Class 5*	*Class 8*	*Class 11*	*Class 14*
Day 2	73.19	2.41	4.22	15.36	4.82
Day 5	4	64	20	2.15	9.85
Day 8	4.19	18.39	62.26	4.84	10.32
Day 11	24.92	1.25	5.61	62.30	5.923
Day 14	4.73	6.80	10.35	2.66	75.45

Table 3. Number of significantly changed features in all pairs of inspection days for set1

Set 1	*Day 2*	*Day 5*	*Day 8*	*Day 11*	*Day 14*
Day 2	0	592.47	503.99	332.48	1288.67
Day 5	592.47	0	4.98	22.94	1118.72
Day 8	503.99	4.98	0	22.01	1025.24
Day 11	332.48	22.94	22.01	0	1026.94
Day 14	1288.67	1118.72	1025.24	1026.94	0

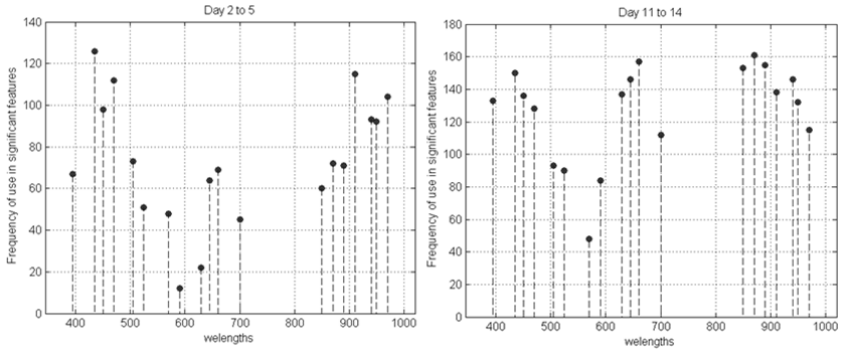


Fig. 5. The Frequency map for contribution of the wavelengths in significant features

In addition, considering the requirements of an industrial level vision system, it is interesting to know which wavelengths contributed most in the significant features. For this reason, we examined the frequency at which a wavelength was included in the features below the critical point. Fig. 5 shows the frequency of each wavelength being used in those features. The frequency maps of the pairwise analysis between days 2 to 5 and 11 to 14 are shown. For day 2 to 5, the three mostly used wavelengths were 435 nm (blue), 910 nm (NIR) and 470 nm (blue). In case of day 11 to 14, the 850-890 nm NIR bands as well as 660 nm (red) and 435 nm (blue) had the highest frequency. Similar analysis for other cases showed that, NIR bands as well as the blue and red wavelengths were among the top frequent bands.

There were some differences between the previous work in [6] and this study. The carrots were fried in oil in that work while no oil was used for this study. The inspection days weren't exactly the same compared to this work and the freezing period was two months more. The elastic-net regression model was built using leave-one-out cross validation that we believe may cause over fitting regarding the limited samples and high number of features. The significant change was found between days 2 and 4 in that work.

5 Conclusion

In this paper, multispectral images of pre-fried carrots were used to detect the changes in their quality within 14 days of inspection. The pre-fried carrots were kept around two months in the freezer and then were moved into the refrigerator. The use of multispectral images helps to extract the features representing the surface color as well as NIR characteristics of carrots. The Benjamin-Hachberg (BH) multiple hypothesis testing method was used to find the most significantly changed features over the storage days. The most important change in carrot samples occurred after 2 weeks. While with less significance level, they also changed after 2 days. Classification results obtained by SVM supported this. However, the elastic-net regression results had high

MSEs. As a result, the 2-sided t-tests on the regression predictions of any set of 2 days at a 5% level were significant.

Acknowledgement:
The authors would like to thank Peter Stubbe (National Food Institute, Technical University of Denmark) and Helene Carlsen (student) for their help in VideometerLab experiments.

This work was (in part) financed by the Centre for Imaging Food Quality project which is funded by the Danish Council for Strategic Research (contract no 09-067039) within the Program Commission on Health, Food and Welfare.

References

1. Adler-Nissen, J.: The Continuous Wok - a New Unit Operation in Industrial Food Processes. J. Food Process Engine. , 25, 435–453 (2002)
2. Adler-Nissen, J.: Continuous Wok-Frying of Vegetables. Process parameters influencing scale up and product quality. J. Food Engineering, 83, 54–60 (2007)
3. Bao, H. N. D., Arason, S., Porarinsdottir, K. A.: Effects of Dry Ice and Superchilling on Quality and Shelf Life of Arctic Charr (Salvelinus Alpinus) Fillets. J. Food Engineering, 3(3), art.7 (2007)
4. Adler-Nissen, J., Akkerman, R., Frosch, S., Grunow, M., Løje, H., Risum, J. Wang, Y., Ørnholt- Johansson, G.: Improving the supply chain and food quality of professionally prepared meals. J. Trends in Food Science & Technology, 29, 74-79 (2013)
5. Løje, H., Dissing, S. B., Clemmensen, H. L., Ersbøll, K. B., Adler-Nissen, J.: Multispectral Imaging of Wok Fried Vegetables In: 18th Scandinavian Workshop on Imaging Food Quality, pp. 59-62. Technical Report, Ystad (2011)
6. Dissing S.B., Clemmensen, H. L., Løje, H., Ersbøll, K. B., Adler-Nissen, J.: Temporal Reflectance Changes in Vegetables. In: 12th IEEE International Workshop on Computer Vision, pp. 1917 - 1922. IEEE press, Kyoto (2009)
7. Clemmensen, H. L., Dissing S.B., Hyldig, G., Løje, H.: Multispectral Imaging of Wok fried vegetables. J. Imaging Science and Technology, 56(2), 20404-1-6 (2012)
8. Dudoit, S., Shaffer, P. J., Boldrick, C. J.: Multiple Hypothesis Testing in Microarray Experiments. J. Statist. Sci. 18, 71-103 (2003)
9. Diz, A. P. , Carvajal-Rodríguez, A., Skibinski, D. O. F.: Multiple Hypothesis Testing in Proteomics: a Strategy for Experimental Work. J. Molecular & Cellular Proteomics, 10, M110.004374. (2011)
10. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2009)
11. Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. J. Royal Statistical Society, 57, 289-300 (1995)
12. Otsu, N.: A threshold selection method from gray-level histograms. J. IEEE Transactions on Systems, Man, and Cybernetics, 9, 62–66 (1979)
13. Gonzalez, R. C., Woods, R. E.: Digital Image Processing. Prentice Hall, New Jersey (2001)
14. Chang, Ch. Ch., Lin, Ch. J.: LIBSVM: a library for support vector machines. J. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, (2011)

Appendix G

Optimal vision system design for characterization of apples using US/VIS/NIR spectroscopy data

Authors: Sara Sharifzadeh[1], Mabel V Martíinez Vega[2], Line H. Clemmensen [1] and Bjarne K. Ersbøll[1].

1. Department of Applied Mathematics and Computer Science, Technical University of Denmark.
2. Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen.

Published in proceedings 20^{th} *International Conference on Systems, Signals and Image Processing (IWSSIP 2013), July 2013, pp.11-14.*

Optimal Vision System Design for Characterization of Apples Using US/VIS/NIR Spectroscopy Data

Sara Sharifzadeh[1], Line H. Clemmensen[3], Bjarne K. Ersbøll[4]
Department of Mathematics and Computer Science
Technical University of Denmark
Copenhagen-Denmark
{sarash, lkhc, bker@dtu.dk}

Mabel V. Martinez Vega[2]
Department of Plant and Environmental Science
University of Copenhagen
Copenhagen-Denmark
mmar@life.ku.dk

***Abstract*— Quality monitoring of the food items by spectroscopy provides information in a large number of wavelengths including highly correlated and redundant information. Although increasing the information, the increase in the number of wavelengths causes the vision set-up to be more complex and expensive. In this paper, three sparse regression methods; lasso, elastic-net and fused lasso are employed for estimation of the chemical and physical characteristics of one apple cultivar using their high dimensional spectroscopic measurements. The use of sparse regression reduces the number of required wavelengths for prediction and thus, simplifies the required vision set-up. It is shown that, considering a tradeoff between the number of selected bands and the corresponding validation performance during the training step can result in a significant reduction in the number of bands at a small price in the test performance. Furthermore, appropriate regression methods for different number of bands and spectrophotometer design are determined.**

Keywords—Sparse regression, spectroscopy, lasso, elastic-net, fused lasso

I. INTRODUCTION

In food industry, the vision-based techniques such as spectroscopic measurements are widely used methods for quality monitoring of the food items. They acquire changes in the chemical and physical composition as factors of quality [1, 2]. For instance, the optical characteristics such as reflectance or absorbance measured by UV/VIS/NIR spectroscopy can represent the pigmentation and structural tissue changes in the plant organs.

There are different types of spectrophotometers used for spectroscopy and their spectral resolution (provided by monochromator) is an important characteristic showing the range of wavelengths they support [1, 3]. However, not all the wavelengths are equally important for characterization of food items. Usually the data in adjacent wavelengths are highly correlated and many of them are redundant, whereas other wavelengths may not carry relevant information for the problem at hand. Therefore, choosing a proper set of wavelengths carrying relevant information will help to simplify the vision system.

This work was financed by the Centre for Imaging Food Quality project which is funded by the Danish Council for Strategic Research (contract no 09- 067039) within the Program Commission on Health, Food and Welfare.

978-1-4799-0944-5/13/$31.00 ©2013 IEEE

The aim of this paper is to solve such problems by employing sparse regression methods on UV/VIS/NIR spectroscopic data (306-1130 nm) of an apple cultivar. Two quality parameters, the sugar content called solvable solid content (SSC) and firmness of the apples [2] were predicted using their spectroscopic data. Sparse regression methods assist to reduce the number of wavelengths [4] and can simplify the vision set-ups used in food quality control [8]. We compared three sparse regression techniques; least angle shrinkage and selection operator (lasso) [4], elastic-net (EN) [4] and fused Lasso (FL) [5]. The data set was divided into different training and test sets four times and the average results are considered. A 10-fold cross validation (CV) was employed for training the prediction models. However, using the model parameters corresponding to the minimum validation error resulted in the use of a considerable number of wavelengths. In order to reduce the number of wavelengths even more, two strategies were investigated in the training phase. First, the one standard error rule was used [4]. In addition, manual selection of the proper number of wavelengths corresponding to an acceptable performance compared to the optimal point was performed. Results showed that both methods reduced the number of wavelengths significantly for all methods. However, this reduction was more considerable for firmness than SSC. In addition, the second strategy decreased the number of required wavelengths more and achieved better performance than the first one. Finally, a relation between the statistical methods and the design of the vision setups were determined.

II. DATA DESCRIPTION

The apple cultivar that was used in this paper is called "Rajka". Spectroscopic measurements were performed on both sides, exposed and non-exposed to the sun, in 825 wavelengths (306-1130 nm) and the average results were considered. There were 185 data points (apple samples) in total. In addition, the SSC (%Brix) and the firmness (N) values for each apple were available from laboratory measurements. Figure 1 shows the spectroscopic data in UV/VIS and NIR wavelengths as well as the corresponding

sorted SSC and firmness signals. A slight trend can be seen in spectroscopic data in both UV/VIS and NIR bands that follows the corresponding sharp changes in SSC and firmness. In order to form the training and test sets, the samples were ranked in ascending order according to the SSC or firmness level. Then, from every 4 fruit samples, one was chosen as test (unseen data during training) and the rest as training. This was repeated 4 times by changing the number of test samples (1,2,3,4) to prepare 4 training and test sets ($X_{tr}, Y_{tr}, X_{ts}, Y_{ts}$).

III. STATISTICAL ANALYSIS

Since the number of samples N=185 was much smaller than the number of wavelengths P=825, the parsimonious regression methods were employed.

A. Lasso

Considering the general regression problem with $(X^i, y_i), i = 1,2,\dots,N$, where $X^i = (x_{i1}, \dots, x_{iP})^T$ are the predictor variables and y_i are the responses, the lasso regression method estimates the regression coefficients β_{lasso} by minimizing the residual sum of squares so that, the L_1 norm of the coefficients is penalized [4]:

$$\hat{\beta}_{lasso} = argmin_{\beta}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{P}|\beta_j|\right\}$$

In $P > N$ case, the lasso selects at most N variables before it saturates [4]. Moreover, it does not support a grouping effect. That means that, if there is a group of variables with high pairwise correlations, then the lasso tends to select only one variable from the group. In this paper, the lasso algorithm implementation based on least angle regression (LAR) algorithm formed in [6] was used. The training involved a 10 fold CV to determine the number of non-zero coefficients varying from 1 to N.

B. EN

EN is a sparse regression method in which the regression coefficients β_{EN} are calculated based on both L_1 and L_2 norms penalty [4]:

$$\hat{\beta}_{EN} = argmin_{\beta}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda_1\sum_{j=1}^{P}|\beta_j| + \lambda_2\sum_{j=1}^{P}|\beta_j|^2\right\}$$

It can be used in ill-posed conditions where N<P. In addition, it has the grouping effect that helps to design a vision system with groups of close wavelengths. The EN implementation from [6] was used in this paper. Training was performed by a 10 fold CV with loops for selection of the

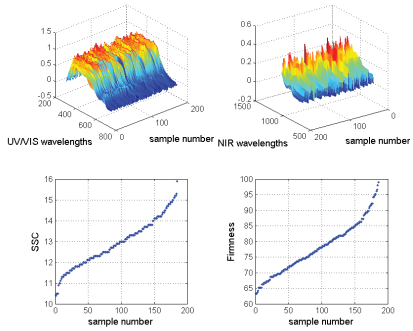


Fig. 1. The UV/VIS and NIR wavelengths sorted spectroscopic data and the corresponding SSC and firmness signals.

norm 2 penalty λ_2 and the number of non-zero coefficients.

C. FL

FL is a generalized version of lasso that encourages sparsity by means of the L_1 norm penalty on both regression coefficients and their successive differences [5]:

$$\hat{\beta}_{FL} = argmin_{\beta}\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda_1\sum_{j=1}^{P}|\beta_j| + \lambda_2\sum_{j=2}^{P}|\beta_j - \beta_{j-1}|\right\}$$

The fused lasso is especially useful for the N<P cases, since it sets many coefficients to zero and finds groups of close features. The first penalty term encourages sparsity in the coefficients and the second one encourages sparsity in their differences. Therefore, with this solution we expect to find groups of adjacent wavelengths. In this paper, the implementation of this algorithm from the SLEP package [7] was used. A 10 fold CV with two loops for the choice of the two penalty terms were used for training the models. In addition, for each trained model the number of bands were calculated in each case and used for making decision about the best model parameters.

D. Model Selection

As mentioned in section 1, the one-standard error rule [4] was used as the first strategy for reducing the number of bands and making the models more parsimonious. The one- standard error rule picks the simplest model within one standard error from the minimum error point. Suppose that, we have P number of variables and M folds. The standard error of the error matrix $err_{P\times M}$ at point p is computed as follows:

$$se_p = \frac{std(err(p,:))}{\sqrt{M}}$$

Computing this value at the minimum point, the best model parameter can be found at one se_p distant from the minimum point in the direction where less bands are chosen. The second selection strategy is simply a manual selection by comparing the error at the minimum point and the points with less number of bands. In fact, a tradeoff was made between the reduction in number of bands and the increase in error.

IV. Experimental Results

In order to evaluate the regression methods the root mean square error and the R-square criteria were used:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}, \quad R^2\% = \left(1 - \frac{RSS}{TSS}\right) \times 100\%,$$

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, TSS = \sum_{i=1}^{n}(y_i - \bar{Y})^2$$

where y_i and $\hat{y}_i$ are the original and estimated response values and $\bar{Y}$ is the mean value of all target values.

In figure 2 the minimum error points as well as the points selected by the one standard error rule and also manual selection are shown for the three methods. Since the dimension of the resulting average validation error map from the 10 fold CV loop varies for the three regression methods, the three plots do not represent an equal number of points. In case of lasso, the number of bands was the only model parameter and thus the average validation error was an N length vector. For EN model, besides the number of bands, the norm 2 penalization coefficient λ_2 was the second parameter. Therefore, the illustration was performed by reducing the dimension into one so that, just the minimum error for each lambda as well as the points found by the two selection strategies are illustrated. Finally, for FL, the two varying parameters were λ_1 and λ_2 that created a 2D average error map. However, in this work we are interested to make the decision based on the number of wavelengths. Therefore, a 2D matrix of the same length showing the average number of wavelengths over the 10 folds was also calculated. Then, these two matrixes were vectorized and shown verses each other. In all of these cases, the selection direction of the new points was defined in a way that the simplest and most parsimonious models could be selected. The density of selection of different wavelengths for the four data sets in SSC estimation is shown in figure. 3. As can be seen, the UV, VIS and NIR regimes are selected by all three methods.

Finally, the average results of the four training and test sets from all the three regression methods for the SSC and firmness are presented in tables 1 and 2. As can be seen, in both cases moving from the minimum point toward a new point selected manually or by one standard error rule significantly reduced the average number of wavelengths N_w

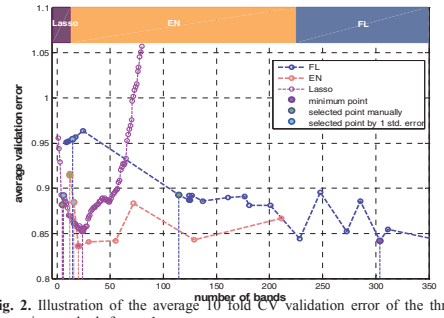


Fig. 2. Illustration of the average 10 fold CV validation error of the three regression methods for set 1.

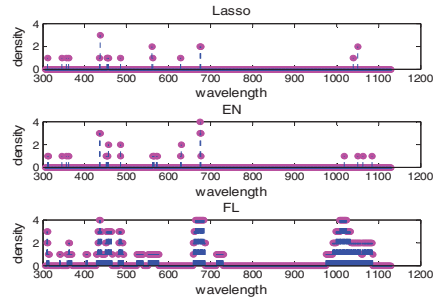


Fig. 3. Density of band selection for SSC estimation using the three regression methods. Model parameters were obtained by manual selection for 4 data sets.

and also reduced the over-fitting by decreasing the difference between the training and test performances. Moreover, the manual selection reduced the number of bands more than the one standard error rule.

A. Discussion

Based on the results, the methods that are suitable for vision set-ups with different number of bands are illustrated on top of figure 2 by different colors. Besides the number of bands, the width of the regimes is important in spectrophotometer design. Considering figure 2 and 3 and the two tables, lasso is suitable when a few individual narrow bands (less than 10 bands) can be provided by e.g. a few LEDs. EN is suitable when more bands (up to 200) in narrow regimes can be supplied. A monochromator capable of selecting a few narrow regimes of laser light suits this case. Finally, FL is the best choice when a lot of bands (e.g. more than 200) in broad regimes of laser light are available. The monochromator does not need to provide high resolution in this case.

Table 1. The average results of the three regression methods for SSC.

SSC		Minimum point	Manual Selection	1 Std. Error Rule
FL	R_{ts}^2%	41.12	39.40	37.80
	R_{tr}^2%	55.70	42.68	43.98
	$rmse_{ts}$	0.86	0.87	0.88
	$rmse_{tr}$	0.76	0.87	0.86
	N_w	532.5	142.75	158.75
EN	R_{ts}^2%	41.28	41.32	37.86
	R_{tr}^2%	53.32	44.92	39.59
	$rmse_{ts}$	0.86	0.86	0.88
	$rmse_{tr}$	0.78	0.85	0.89
	N_w	30.25	11.50	19.25
LASSO	R_{ts}^2%	41.55	40.87	35.16
	R_{tr}^2%	55.46	44.39	37.02
	$rmse_{ts}$	0.86	0.86	0.90
	$rmse_{tr}$	0.76	0.85	0.91
	N_w	16.0	6.25	5.0

Table 2. The average results of the three regression methods for firmness.

Firmness		Minimum point	Manual Selection	1 Std. Error Rule
FL	R_{ts}^2%	47.33	41.53	36.52
	R_{tr}^2%	67.41	42.89	46.35
	$rmse_{ts}$	5.74	6.05	6.29
	$rmse_{tr}$	4.71	6.24	5.99
	N_w	795.0	22.25	179.0
EN	R_{ts}^2%	47.50	39.96	44.75
	R_{tr}^2%	68.46	39.87	53.72
	$rmse_{ts}$	5.73	6.14	5.88
	$rmse_{tr}$	4.63	6.39	5.60
	N_w	758.5	16.0	307.5
LASSO	R_{ts}^2%	45.17	41.97	33.20
	R_{tr}^2%	68.87	44.12	36.43
	$rmse_{ts}$	5.86	6.03	6.47
	$rmse_{tr}$	4.60	6.17	6.56
	N_w	36.75	4.0	4.75

I. CONCLUSION

In this paper, three regression methods; lasso, EN and fused lasso were used for estimation of the SSC and firmness level of an apple cultivar using their spectroscopic data. By manual selection of a new point or using the one standard error rule instead of the minimum error point, we could significantly reduce the number of required wavelengths for training the prediction model at a price of a small increase in the test error. Finally, the proper regression methods with different number of bands and types of spectrophotometer design were defined in the discussion.

REFERENCES

[1] B. Herold, S. Kawano, B. Sumpf, P. Tillmann, and K. B. Walsh, *Book Chapter: VIS/NIR spectroscopy,* CRC Press, London, 2009.

[2] B. Nicolai, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K. Theron and J. Lammertyn, " Non-destructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review," *Postharvest Biology and Technology*, Elsevier, pp. 99-118 , 2007.

[3] D. Wen Sun, *Hyperspectral imaging for food quality analysis and control*, Elsevier academic press, United Kingdom, 2010.

[4] T.Hastie, R. Tibshirani and F. Jerome, *The Elements of Statistical Learning*, Springer, New York, 2008.

[5] R. Tibshirani and M. Saunders, "Sparsity and smoothness via the fused lasso," *Journal of royal statistical society* (67), Blackwell Publishing, United Kingdom, pp. 91–108, 2005.

[6] K. Sjöstrand, "Regularized Statistical Analysis of Anatomy," *PhD thesis*, Department of Informatics and Mathematical Modeling, Technical University of Denmark, 2007.

[7] J. Liu, S. Ji, and J. Ye. "SLEP: Sparse Learning with Efficient Projections," Arizona State University, 2009.

[8] S. Sharifzadeh, J. L. Skytte, O. H. A. Nielsen, B. K. Ersbøll, L. K. H. Clemmensen, "Regression and Sparse Regression Methods for Viscosity Estimation of Acid Milk From it's SLS Features," Proceedings: IWSSIP 2012, 11-13 April 2012, pp. 58-61, Vienna, 2012.

Appendix H

Sensory Quality Prediction Using Multispectral Imaging

Authors: Sara Sharifzadeh[1], Hanne Løje[2], Grethe Hyldig[2], Line H. Clemmensen[1], Bjarne K. Ersbøll[1].

1. Department of Applied Mathematics and Computer Science, Technical University of Denmark.
2. National Food Institute, Technical University of Denmark.

Submitted.

Sensory Quality Prediction Using Multispectral Imaging

Sara Sharifzadeh [1*], Hanne Løje[2], Line H. Clemmensen[1], Grethe Hyldig[2], Bjarne K. Ersbøll[1]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, DK-2800 Lyngby, Denmark {sarash, lkhc, bker}@ dtu.dk, (Tel: +45- 45 25 53 51)

[2]National Food Institute of Denmark, Technical University of Denmark, DK-2800 Lyngby, Denmark {halo, grhy}@food.dtu.dk

(The first two authors contributed equally to this work.)

Abstract

The use of computer – vision based systems as non-destructive and in-line quality monitoring methods in food industry is increasing. We propose the use of multispectral images ranging from visible (VIS) to near infrared bands (NIR[1]) to predict sensory attributes. Sensory evaluation is an important quality assessment method in food industry. However, it is time consuming and in some cases destructive. On the other hand, multispectral imaging is a non-invasive and cheap method that can be used fast and in-line. The visible spectra show the pigmentation and appearance information while the NIR spectra are correlated to the chemical characteristics of the object under study. In this work, two types of vegetables (carrot and celeriac) were used for investigations. Two batches of stir-fried vegetables were evaluated after a freezing period followed by a chill-storage period for up to 14 days at 5°C. At each day of experiment, the sensory evaluation was performed by a sensory panel of 6 assessors. In addition, multispectral images were acquired from the same samples in 19 different wavelengths (VIS-NIR). The aim is to explore the general relationship between the sensory attributes and the multispectral images. We develop statistical regression models to predict the sensory attributes from the spectral information. Experimental results demonstrated such a relationship between some of the sensory attributes and spectral features. From the obtained results, we found that variation of sensory scores over the days of storage and their consistency over the batches were the two main requirements for generalization of the models. We also found that, due to the limitation in human visual perception, the color attribute did not have such characteristics. In addition, the analysis results demonstrated that both visible as well as NIR wavelengths were among the most contributing wavelengths in the models.

VIS-NIR: visible-near infrared
CCD: Charge coupled device
QIM: quality index method
EN: elastic-net
RMSE: root mean square error
SSC: solvable solid content
PLSR: partial least square regression
DPA: Discrimination Power Analysis
DP: Discrimination Power

Key words (6):
Multispectral imaging, Sensory assessment, Stir-fried vegetables, Sensory attributes, Regression

1 Introduction

The use of computer vision - based systems for assessment and monitoring the quality of food items has gained attention widely in recent years. A simple example is the use of a digital CCD camera for quality inspection of apples (Garrido-Novell, et al., 2012).

The classic methods for food quality assessment are mainly based on laboratory tests and sensory evaluation, usually performed by human experts. However, such methods have some limitations. For example they can be destructive in some cases and they are dependent on well trained assessors. In addition, they are slow methods when used in-line in a production line.

Due to these limitations, the computer vision - based techniques such as multispectral imaging have been employed as an alternative for quality inspection of food items. These techniques are fast, non-invasive and result in reproductive quality monitoring methods in food industry. Additionally, they can be used objectively and in-line. Multispectral imaging gives information about the color and visual characteristics of the food under study as well as its chemical characteristics that are correlated to its quality (ElMasry & Sun, 2010). That is based on certain materials unique spectral signatures in the electromagnetic spectrum (Sun , 2009). Such spectral imaging systems can be designed very cheap for food quality monitoring.

In most cases the assessment is performed by detection or prediction of a "quality parameter" such as appearance condition (color or texture) or content level (sugar, acidity, etc.). Reviewing the literature shows that there are only a few research works on the use of vision-based systems for prediction of the human attitude about the food quality which we call "sensory attribute" or "sensory score" in this paper.

Sensory analysis is one of the important methods for evaluation of the eating quality of food items and consumer satisfaction in food industry. Usually a panel of well-trained experts or untrained consumers evaluates a food product. There are several qualitative or quantitative sensory evaluation methods (Varela & Ares, 2012; Lawless & Heymann, 1999) . However, sensory analysis in some cases is a destructive method and cannot be used inline. In addition, it can be very expensive and time consuming. Therefore, it cannot be used as a routine analysis in an industrial production and processing line (Kamruzzaman , ElMasry , Sun, & Allen, 2013)

This paper addresses the problem of prediction of sensory attributes of wok-fried vegetables, (carrot and celeriac) using multispectral imaging techniques. Such kind of research for other types of food items were conducted before. Prediction of sensory attributes related to the eating quality of lamb meat samples using VIS-NIR spectroscopy (400-2498 nm) (Andrés, et al., 2007) , lamb meat tenderness using NIR hyperspectral images (900-1700 nm) (Kamruzzaman , ElMasry , Sun, & Allen, 2013), pork samples using NIR hyperspectral images (900-1700 nm) (Barbina, ElMasrya , & Sun, 2012), table grapes by hyperspectral images in VIS-NIR range (400-1000 nm) (Baiano, Terracone, Peri, & Romaniello, Application of hyperspectral imaging for prediction of physico-chemical and sensory characteristics of table grapes, 2012) are the main previous works in this case.

In this work, we study the relationship between multispectral images and sensory attributes of vegetables. The wok-fried vegetables (carrot and celeriac) used in this work, were stored in refrigerator for some days after a freezing period (Adler-Nissen, et al., Improving the Supply

Chain and Food Quality of Professionally Prepared Meals, 2013). Previously, multispectral images of the same type of vegetables were analyzed to find significant changes over a storage period and the results showed that multispectral imaging is capable to distinguish the subtle visual changes of samples over the storage days (Sharifzadeh, Clemmensen, Løje, & Ersbøll, 2013; Dissing, Clemmensen, Løje, Ersbøll, & Adler Nissen, 2009; Clemmensen, Dissing, Hyldig, & Løje, 2012). Therefore, in this work, multispectral images are used to predict a wide range of sensory attributes including color, taste, smell and texture.

The most similar work for vegetables was reported in (Løkke, Seefeldt, Skov, & Edelenbos, 2013) that also has differences in both type and condition to our work. In that work, the relationship between the multispectral images (405-970 nm) of green vegetables (wild rocket) and two sensory attributes (color and texture quality) was studied in changing condition of storage temperature, time and packaging condition.

The objective of this study is first to investigate, if there exists a general relationship between the sensory attributes and multispectral images. This can be tested by developing statistical regression models for prediction of sensory attributes using multispectral images of vegetables. Secondly, in the case of such relationship, to find "strategies" or "factors" and possible "requirements" for improving the predictive regression models with general prediction ability on new batches of data. This might be possible using a combination of sensory attributes or the specific individual attributes. Finally, it is to determine the spectral wavelengths carrying the most relevant information regarding the prediction. Finding such wavelengths is important for the design of an appropriate imaging system in industry for measuring relevant quality parameters.

2 Material and methods

2.1 *Material preparation*

Two types of vegetables, carrot and celeriac are selected in this work. They are prepared based on a new way of producing convenience vegetable products of high culinary quality by continuously stir-frying the vegetables (Adler-Nissen J. , 2007). Stir-fried root crops produced by the continuous stir-frying process have a robustness against freezing and exhibit no visible drip losses or only little drip loss after thawing (Adler-Nissen J. , 2007; Clemmensen, Dissing, Hyldig, & Løje, 2012) compared to some blanched and frozen vegetables.

For this aim, the raw celeriac and carrots were cut into cubes of size approximately 0.5 cm cubed. Due to the relatively high biological variations of vegetable products, we used two batches for each type of vegetables (B_1, B_2). This helps to include more possible variations of the original population into the data set, that helps for generalization of the statistical prediction models. Batch one was harvested and pre-processed (washed and cut) on one day, and the other batch on the following day. Additionally, there were two replicate samples in each batch (a,b), in order to have estimates of both within batch variations as well as between batch variations. Covering both these variations is important for developing a general prediction model for sensory attributes. In a pilot plant, the raw vegetables were stir-fried using a special frying machine;" the continuous wok" (Adler-Nissen, 2007). After frying and cooling, the products were packed in polyethylene bags in 500g portions and frozen to −30˚C. After about 60 days of freezing, the bags with the stir-fried vegetables were removed from the freezer and thawed up to 14 days at +5˚C in a refrigerator.

2.2 *Acquiring multispectral images*

On each day of analysis (days 2, 5, 8, 11 and 14), two polyethylene bags of the two batches (B_1 , B_2) were taken out of the refrigerator. Then, for each member of the sensory panel, 30 g of each sample bag was placed in two petri dishes (a,b). To acquire the multispectral images, a VideometerLab was used (Carstensen, Hansen, Lassen, & Hansen, 2006) and each petry dish was digitized separately (see Figure 1). VideometerLab is a multispectral imaging device designed for fast and accurate determination of surface color and chemical composition. It was used and described in detail in (Dissing, Clemmensen, Løje, Ersbøll, & Adler Nissen, 2009). The multispectral images were captured at 19 different wavelengths ranging from 430 to 970 nm.

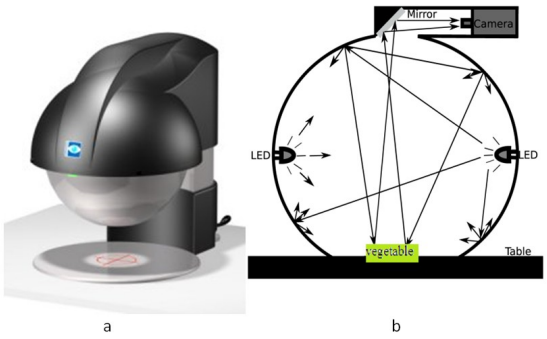


Figure 1. (a) a VideometerLab (b) internal design of a VideometerLab (Dissing, Nielsen, Ersbøll, & Frosch, 188 2011)

2.3 *Sensory evaluation*

An internal panel consisting of 6 assessors performed the evaluation. They were all selected and tested according to international standards (ISO-8586-1, 1993) for their ability to make sensory evaluations. The sensory score was developed based on the agreement of the 6 assessors about each sensory attribute on each vegetable sample.

At each day of analysis, after measurement with the VideometerLab, the petry dishes (a,b) of each batch or sample bag (30 g) were transferred to an aluminum tray (one aluminum tray for each sample) and re-heated to be served for the assessors. Then, for each assessor at each analysis day, two replicates (a,b) of each batch (B_1, B_2) were served (4 petry in total).

In this work, the sensory attributes were analyzed using a sensory method named Quality Index Method (QIM). In another study (Clemmensen, Dissing, Hyldig, & Løje, 2012), individual QIM schemes for carrots and celeriac were designed in analogy to how QIM scheme was developed for several fish species (Martinsdóttir, Schelvis, Hyldig, & Sveinsdóttir, 2009). The sensory attributes were defined based on appearance, smell, taste and texture categories. For both vegetable types 7 sensory attributes were defined. For carrot, the attributes were discoloration, smell, cloying sweetness, taste, frying aroma, off-taste and firmness and for celeriac they were discoloration, smell, frying aroma, taste, sweetness, off-taste and firmness. Each attribute was

given a score between zero and two or three demerit points, so that the score increases as quality decreases, i.e. very fresh products have scores near to 0. More information about QIM are provided in the Appendix.

2.4 *Multispectral image analysis*

Spectral images of a sample petri dish of carrots are shown in Figure 2. At the first step of image analysis, the vegetable pieces were segmented from the background and from each other using image segmentation techniques such as thresholding, morphological operation and filtering (Gonzalez & Woods, 2001; Otsu, 1979). Each vegetable piece was considered as an individual sample. Then, a feature vector was formed for each detected vegetable piece. First, for each of the 9 bands, all its 18 possible ratios to the other bands were calculated (totally 342 ratio images). The band ratioing method is one of the simplest methods for multispectral image enhancement technique (Jain, 1989). It is usually applied to enhance the spectral differences between raw images and suppress the effect of undesired effects such as variable illumination and slop shadows. Then, in each ratio image, 9 percentiles (1, 5, 10, 25, 50, 75, 90, 95 and 99) were calculated for each piece. In this way, feature vectors are formed from the spectral images with the length of $P = 3078\ (9 \times 342)$ per piece of vegetable. In the rest of the paper we refer to them as "spectral features".

2.5 *Statistical analysis*

In order to develop prediction models for sensory attributes using the spectral features, there must be significant changes or variations in both spectral features and sensory attributes. The statistical test used for existence of such significant variations in spectral features over the days

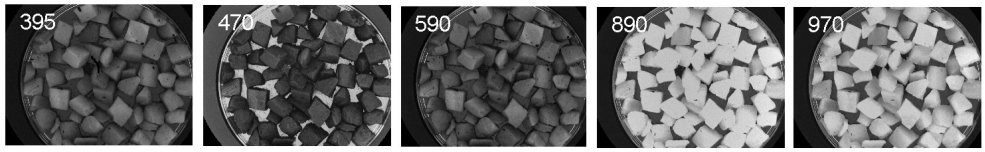


Figure 2. Spectral images of carrot in some wavelengths, shown in Nano-meter range.

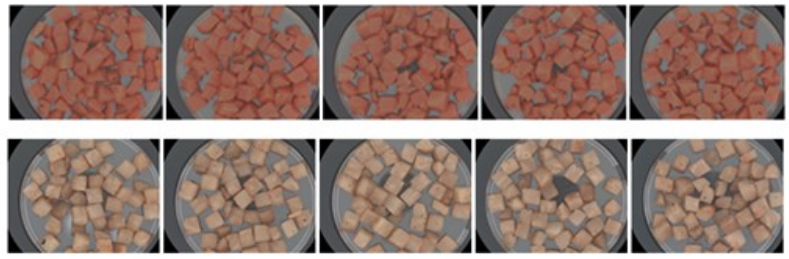


Figure 3. Pseudo-RGB images of carrot (top) and celeriac (bottom). From left to right: day 2, 5, 8, 11and 14.

of storage will be presented in the following and the variation of the sensory attributes will be shown only by visualization and due to the considerations about the length of the paper, the

statistical tests about the significant changes of the sensory data will not be presented. Visualing the sensory attributes also helps to obtain better insight about the variation of the sensory data as will be explained in the following.

The variation analysis gives information about the changes in freshness of the vegetables over the storage days. However, the aim of this work is not to test the freshness of vegetables. We re-emphasize that the existence of variation in both the spectral features and sensory attributes is important for developing the prediction models.

2.5.1 Statistical analysis of spectral features

It is not possible to distinguish any differance between the vegetable samples over the storage days visually. This can be observed from the pseudo-RGB images of carrot and celeriac samples for five inspection days shown in Figure 3. However, an analysis of the multispectral images demonstrates significant variations over the storage days (Dissing, Clemmensen, Løje, Ersbøll, & Adler Nissen, 2009; Sharifzadeh, Clemmensen, Løje, & Ersbøll, 2013) . The spectral features are used in a statistical analysis based on multiple hypothesis testing (Hastie, Tibshirani, & Friedman, 2009; Benjamini & Hochberg, 1995). More information about this test is provided in the appendix. The test is performed on the spectral features of all possible pairs of days. The output of each test is the number of significantly changed features between the two days. To make a decision about a significant change between two days of storage, there must be high number of significant features between all pairs of days between them. For example, it is necessary to observe a high number of significant features between both days (2,5) and (5,8) to consider a significant change between days (2,8). The result of this analysis for carrot was presented in (Sharifzadeh, Clemmensen, Løje, & Ersbøll, 2013).

2.5.2 Visualisation of the sensory attributes

Figure 4 and Figure 5 show the scores of the sensory attributes for the two batches of vegetables. As can be seen some of the sensory attributes have variation over the days. Besides that, two main issues can be observed; the scores from the two batches are different in most cases. In addition, there is not a consistent trend from lower scores toward the higher ones (freshness toward lower qualities), as the number of storage days increases. In some cases such as day 14 of the first batch of carrot for smell attribute, there is even inconsistency between the two replicates of the same batch. The reason can be due to the variability in different pieces of the same vegetable, between the two batches or the panelist's perceptions. Generally, the types food products used in this study have a large variation.

2.6 *Prediction of Sensory attributes using spectral features*

Considering the simplest form of a linear prediction model:

$$\hat{Y} = X\beta + \epsilon \qquad 2.1$$

To predict the sensory attributes using the spectral features formed from multispectral images, each sensory attribute (its sensory scores) are considered as the response vector $\hat{Y}$, $\beta_{P\times 1}$ is the vector of regression parameters, ϵ is the error and the input matrix $X_{n\times P}$ is the matrix of spectral features of all samples (petry dishes). To form the X matrix, the median of all vegetable pieces inside a petri dish was considered as a unique sample in each row of X.

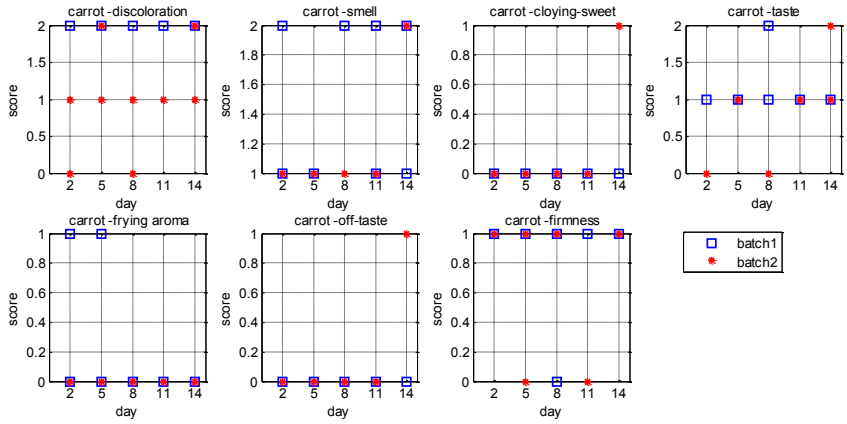


Figure 4. Comparison of the sensory scores for carrot attributes. There are two replicates per batch for each attribute at each day. The replicate samples are overlaid in most cases.

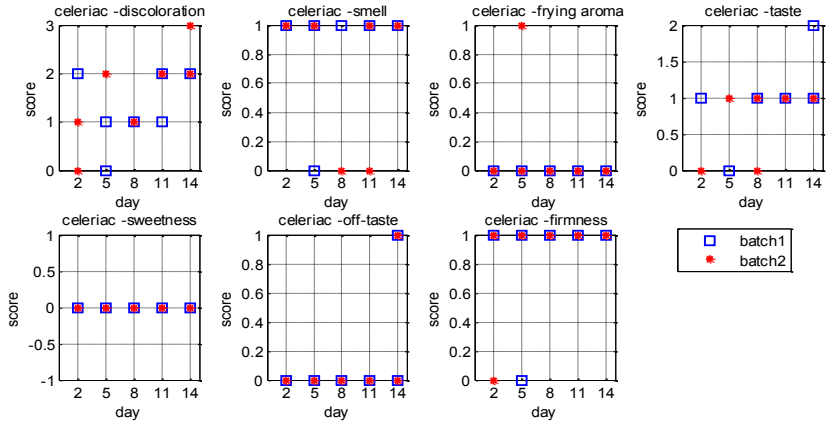


Figure 5. Comparison of the sensory scores for celeriac attributes. There are two replicates per batch for each attribute at each day. The replicate samples are overlaid in most cases

2.6.1 Pre-processing of the matrix of spectral features (X)

The number of columns in X is equal to the number of spectral features (3078) and compared to its rows that show the number of samples is very high. This makes it difficult to train the prediction models. To alleviate this problem, a "pre-processing" was performed on the spectral features to reduce their dimensionality. For this aim, the significant features, found between pairs of days in section 2.5.1 were considered. Then for each of the 3078 features, the number of times that it was significant in the analysis of pairs of days was counted and a vector of counts was formed $Vec_{1\times3078}$. Finally those spectral features that have been significant in most of the tests (have high values in Vec) were selected to be used for developing the prediction models.

2.6.2 Pre-processing of the sensory attributes (Y)

As mentioned before, suitable sensory attributes for prediction models should have some variation over the storage days. However, as shown in Figure 4 and Figure 5, some of the sensory attributes do not show temporal variation during the days of experiments in both batches. These attributes aren't considered for the analysis and the remaining varying attributes (active attributes) of each batch (B_1, B_2) are used. The active attributes are not completely similar for the two batches.

2.6.3 Regression method

The EN linear regression method (Zou & Hastie, 2005) is used for building the prediction models. Compared to the other linear and non-linear regression methods, this method obtained better results. EN is a sparse regression method. This means that some of its regression coefficients β are zero. The sparsity is obtained by penalization of the EN regression coefficients as follows:

$$\hat{\beta}_{EN} = argmin_{\beta}\left\{\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda_1 \sum_{j=1}^{P}|\beta_j| + \lambda_2 \sum_{j=1}^{P}|\beta_j|^2\right\} \qquad 2.2$$

where P is the number of used spectral variables and n is the number of observations (sensory scores or tested petry dishes). The norm one (l_1) part of the penalty generates a sparse model (zero coefficients) and the quadratic part of the penalty removes the limitation on the number of selected variables and encourages a grouping effect. Therefore, it can cancel out the noise effect. In addition, EN is an appropriate method when the ($P \gg n$) which is called an 'ill-posed' problem (Hastie, Tibshirani, & Friedman, 2009) that is the case in our work.

In order to evaluate the accuracy of the prediction models, the RMSE is used.

2.6.4 Prediction tests

Based on the objectives of this work, three different prediction tests are performed.

2.6.4.1 Test I

The first objective of this work is to investigate the general relationship between the sensory attributes and multispectral images. To address this, regression models are developed for prediction of each batch sensory attributes using the spectral features of the same batch. This will help to evaluate the predictability of each batch data.

Another objective is to find the important factors for improving the prediction ability. One factor that influences the performance of a regression problem is to have a response vector Y that has maximum variation especially in accordance to the variation of the input matrix X. In our work, we expect to observe the most important variation in the spectral features (X) and the sensory attributes (Y) between the days of storage. As a first strategy, we improve the response vector Y in terms of variation and discrimination between the days of storage. As mentioned in section 2.6.2, the sensory attributes that show some variation over the storage days in each batch are already moved to an active set of attributes $\{a_1, a_2, \dots, a_k\}$. To improve the variation further, some of the active attributes are combined to form a response vector Y with maximum variation and discrimination between the days of storage. The selection is performed based on the DPA, introduced in (Dabbaghchian, Ghaemmaghami, & Aghagolza, 2010). Considering each day of storage as a class (5 classes), we have chosen those sensory attributes of the active set $\{a_1, a_2, \dots, a_k\}$ that their combination result a_{comb} has the highest DP. A DP is the ratio of "between class variance (S_B)" to "within class variance (S_W)". In fact, we have selected those attributes that when summed, the result

of their summation a_{comb} has scores that are the closest to each other in each day of analysis (class) and the furthest from the scores of the other days (classes). The computation is as follows:

$$S_B = \sum_{i=1}^{5} n_i (m_i - m_t)(m_i - m_t)^T \quad 2.4$$

$$S_W = \sum_{i=1}^{5} \sum_{j=1}^{n_i} (x_j - m_i)(x_j - m_i)^T \quad 2.3$$

$$DP_{a_{comb}} = \frac{S_B}{S_W} \quad 2.5$$

$$Y = \max\{a_{comb(1)}, a_{comb(2)}, \dots, a_{comb(c)}\} \quad 2.6$$

where n_i is the number of scores of a_{comb} in each day (class), m_t is the total average of the a_{comb} scores in all 5 days, m_i is the average of the a_{comb} scores in each of the five storage days and c is the number of tested combination of attributes.

Then, this new response vector Y is used as the response variable for developing the prediction model.

Another important factor to improve the prediction models is to form appropriate calibration and validation sets so that, they cover the existing variation of the batch samples. For this aim and as the second improvement strategy, a systematic sampling method called a "smooth arrangement" or "smooth fractionators" is employed in this work (Gundersen, 2002). The smooth arrangement is formed by ranking the sensory scores in the response vector Y in increasing order. Then, every second score was pushed out from the order and moved to its end so that a monotonically increasing and then decreasing ordering of scores are formed. From this new ordering, a predefined systematic sampling interval of '4' was applied to obtain approximately 25% of the samples for the test set. The remaining 75% of the samples comprised the training set. Both calibration and validation sets comprise the original variation of the data by using this method. The calibration and validation sets are formed four times using this method. The average results are used to evaluate the general performance of the regression method on the data sets. This strategy ensures that we represent all the variation in the data in our calibration set. However, the use of the same batch data for both calibration and validation in the first test introduces some uncertainty to whether the results will generalize to unseen batches. For this reason, the second test was performed to address the next objective of this paper about the generalization.

2.6.4.2 Test II

In this test, one batch was totally used to calibrate a prediction model and the other batch data was used for validation and vice versa. That is a more general test to evaluate the prediction models on totally new samples, i.e. we can test if the models will generalize to unseen batches. For the response vector Y, the same combination strategy used in the test I was used. However, we are also interested to find the individual sensory attributes that have possible links to the spectral features as one of the objectives of this paper. This is performed in the final test.

2.6.4.3 Test III

In the third test, the second analysis (test II) was repeated on individual attributes instead of their combination. The aim is to find the most effective attributes on the prediction models. Then more than one prediction model is built in this test. This analysis is performed only on the active

set of attributes of each batch that have some variation along the storage days $Y = a_i, i = \{1,2, \dots, k\}$. For example, in the case of the first batch of carrots shown in Figure 4, only the smell, taste, frying aroma and firmness can be used for this experiment.

2.7 *Analysis of significant wavelengths*

As the last objective of this paper, the most important wavelengths are found in two steps; first, the wavelengths contributing to the significant spectral features (X) found from multiple hypothesis testing between pairs of days (explained in section 2.6) are considered (f_1). In the next step, the vector of EN regression coefficients β_{EN} is used. Since EN is a sparse method, the irrelevant and noisy spectral feature's coefficients are zero in β_{EN}. Therefore, among the features of f_1, those with non-zero elements in β_{EN} are found (f_2). They are correlated to the sensory response (Y). Each spectral feature was built based on the ratio of two spectral wavelengths (section 2.4). Therefore, both contributing wavelengths in each of the features of (f_2) carry relevant information about their corresponding sensory response. Thus, for each wavelength, the number of times that it is involved in the formation of f_2 features is computed and a density map (D) is formed for all of the wavelengths that can be used for finding the most important wavelengths.

3 **Experimental results**

In this section, the results of analyses explained in the previous section will be presented.

3.1 *Multispectral image analysis results*

Based on the number of significantly changed spectral features, the days of significant change was found. The results are shown in Table 1. These results show that there are significant variations in the features extracted from the multispectral images. It also shows that the spectral images of the two batches of the same vegetables are not exactly similar.

Table 1. The days of significant change found by the analysis of spectral images

	Batch 1	Batch 2
Carrot	2,14	2,8
Celeriac	2,5,8,11,14	8,14

3.2 *Prediction results*

In evaluation of the prediction models, the RMSE was compared with the standard deviation of the corresponding population.

For the prediction result of test I, the average and standard deviation of the RMSE from the four calibrations and validation sets formed by the smooth arrangement are presented in Table 2. The standard deviations of the response vectors are also shown. They are comparable with the RMSEs. The last row shows the active attributes contributing in the combined sensory response vector. In Figure 6, the predicted response vector (Y_h) versus the real response vector (Y) for the validation sets are shown. The pink line shows the ideal case.

The prediction test II was based on training the models using one of the B_1 or B_2 as the calibration batch (B_{cal}) and validating using the other batch (B_{val}). The results of this test are shown in Table 3. These results show the difficulty of prediction for a new batch of sensory data

using a pre-trained model. From a statistical point of view, this may be due to training the model based on the response vectors Y formed from the active attributes of the calibration batch (B_{cal}) , while such attributes are not necessarily active in the validation batch (B_{val}). This can be seen in the last row of Table 3. For example, in the last column, the celeriac model is trained based on the active attributes of the second batch (B_2), smell, frying aroma and off taste, while in the validation batch (B_1) just two of them (smell and frying aroma) together with discoloration, taste and firmness are the active attributes. This inconsistency in the two batches may be due to their limited number of samples that do not capture all possible variation of the population. Naturally, there is variation between the vegetable pieces even when they come from the same batch (within batch variation). Besides that, there is also a biological variation between the two batches (between-batch variation). When the level of variation is high, more samples are required for forming a general prediction model.

In the test III, the single attributes with some variation over the days of storage were used for training the prediction models. Table 4 and Table 5 show the results of this test for carrot and celeriac respectively. Training the models based on the single attributes makes the prediction results more comparable between the calibration and validation sets. This also helps to find those single attributes appropriately correlated to the spectral data. For carrot, the smell and taste attributes obtained the best results with consistency between the two batches. These two attributes have some variation over the days in Figure 4 and some of their two batches scores are similar. In the case of celeriac, the off-taste attribute gave the most consistent result between the two batches. It does not have a lot of variation between the days as shown in Figure 5. However, it is completely consistent between the two batches. After that, smell gave the second best result. It has more variability and less consistency than the off-taste scores. The results of smell and off-taste are plotted in Figure 7 for carrot and celeriac respectively.

Table 2. Results of the prediction test I. The active attributes were combined and the prediction was performed on each batch sensory data using the image features of the same batch.

Method: EN	Carrot-$\boldsymbol{B_1}$	Carrot-$\boldsymbol{B_2}$	Celeriac-$\boldsymbol{B_1}$	Celeriac-$\boldsymbol{B_2}$
Calibration RMSE	0.42±0.06	0.92±0.16	0.57±0.18	0.36±0.08
Validation RMSE	0.48±0.04	1.17±0.07	1.19±0.57	0.49±0.04
Population std.	0.50	1.71	2.27	0.82
Used Attributes	Smell	Smell, Cloying Sweet, Off-taste, firmness	Discoloration, Smell, Taste, Off-taste, Firmness	Smell, Frying Aroma, Off-Taste

Table 3. Results of the prediction test II. The active attributes were combined and one batch was totally used for calibration and the other batch data was used for validation and vice versa.

Method: EN	Carrot		Celeriac	
	$\boldsymbol{B_1}$ cal. $\boldsymbol{B_2}$ val.	$\boldsymbol{B_2}$ cal. $\boldsymbol{B_1}$ val.	$\boldsymbol{B_1}$ cal. $\boldsymbol{B_2}$ val.	$\boldsymbol{B_2}$ cal. $\boldsymbol{B_1}$ val.
Cal. RMSE	0.46	0.32	0.38	0.32
Val. RMSE	0.50	1.89	2.68	3.89
Used Attributes for training	Smell	Smell, Cloying Sweet, Off-taste, firmness	Discoloration, Smell, Taste, Off-taste, Firmness	Smell, Frying Aroma, Off-Taste

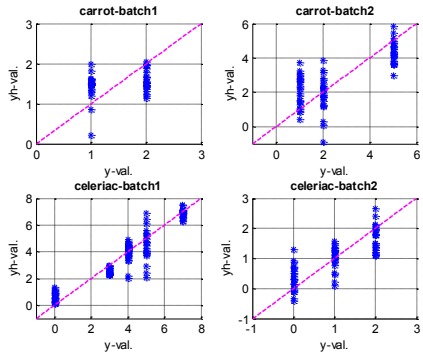


Figure 6. The predicted response vector (Y_h) versus the real response vector (Y) for validation sets of the test I.

Table 4. Results of the third prediction test for carrot. Similar to the second test, one batch was totally used for calibration and the other batch data was used for validation and vice versa. However, individual attributes were used instead of their combination to find the most effective attributes for prediction.

Carrot		discoloration	smell	cloying sweet	taste	frying aroma	off-taste	firmness
$\boldsymbol{B_1}$ cal. $\boldsymbol{B_2}$ val.	Cal. RMSE	-	0.34	-	0.28	0.24	-	0.05
	Val. RMSE	-	0.71	-	0.63	1.02	-	0.57
	pop. Std.	-	0.50	-	0.50	0.50	-	0.50
$\boldsymbol{B_2}$ cal. $\boldsymbol{B_1}$ val.	Cal. RMSE	0.22	0.17	-	0.19	0.29	0.14	0.27
	Val. RMSE	1.43	0.54	-	0.42	1.58	0.56	1.06
	pop. Std.	0.82	0.50	-	0.50	0.82	0.50	0.50

Table 5. Results of the third prediction test for celeriac. Similar to the second test, one batch was totally used for calibration and the other batch data was used for validation and vice versa. However, individual attributes were used instead of their combination to find the most effective attributes for prediction.

celeriac		discoloration	smell	Frying aroma	taste	Sweetness	off-taste	firmness
$\boldsymbol{B_1}$ cal. $\boldsymbol{B_2}$ val.	Cal. RMSE	0.24	0.22	-	0.05	-	0.00	0.15
	Val. RMSE	1.14	0.80	-	0.75	-	0.32	0.60
	Pop. Std.	0.82	0.50	-	0.82	-	0.50	0.50
$\boldsymbol{B_2}$ cal. $\boldsymbol{B_1}$ val.	Cal. RMSE	0.26	0.26	0.19	0.27	-	0.02	-
	Val. RMSE	3.75	1.09	1.43	2.07	-	0.32	-
	Pop. Std.	1.12	0.50	0.50	0.50	-	0.50	-

3.3 *Wavelength analysis results*

As explained in section 2.7, we are interested in finding the wavelengths most correlated to the vegetables quality. Therefore, the density map (D) obtained in section 2.7 is visualized for the first batch of carrot and celeriac in Figure 8. As can be seen, some of the visible band wavelengths around the yellow and orange color (630 and 645 nm) as well as some of the NIR bands are among the most frequently contributing wavelengths. The results of this analysis for the second batch of these vegetables were similar to the first batch and are not presented here. The visible wavelengths are mostly correlated to the pigmentation or color characteristics of the

vegetables, while the NIR wavelengths are mostly correlated to their chemical characteristics (Herold, Kawano, Sumpf, Tillmann, & Walsh, 2008).

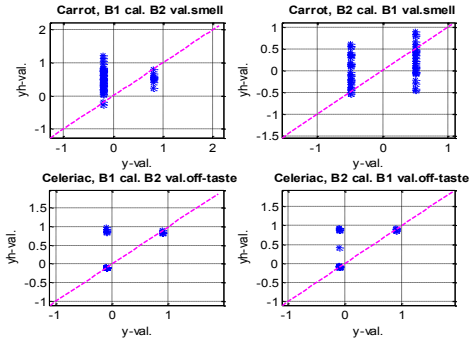


Figure 7. Illustration of two predicted sensory attributes versus their corresponding real values from the third test. For carrot, smell and for celeriac, taste attribute obtained the best result in this test.

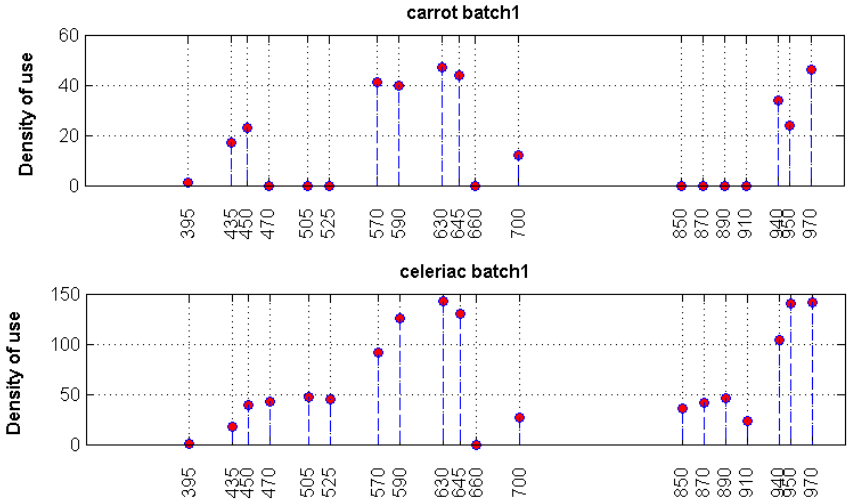


Figure 8. The wavelength analysis results for the first batches of vegetable.

4 Discussion and future work

The results of the first prediction test confirm that there is a link between the sensory attributes and spectral features extracted from the multispectral images.

However, the second test result showed that, it is difficult to build a general prediction model using limited calibration samples. This matches to the research findings of similar works on other food items (Barbina, ElMasrya , & Sun, 2012; Baiano, Terracone, Peri, & Romaniello,

2012). The reasons can be described based on two main issues; first is the wide variability in the population (with-in and between batches) that is not captured in the limited calibration samples. The second reason is the inconsistency in variations of the active attributes of the two batches.

Finally, the analysis of the single attributes made it possible to distinguish the attributes that fulfilled these two issues. So that, the resulting prediction models have an acceptable level of generalization linking the spectral data to the corresponding quality attributes such as off-taste and smell. Then, variability and consistent variation can be considered as the requirements of generalization which was another objective of this work.

On the other hand, the analysis of the wavelengths helped to distinguish the most contributing bands in the significant features used in the prediction models. There were some of the NIR bands among the most frequently used wavelengths. This can explain the reason for the absence of discoloration among those successful single attributes. The discoloration is mostly related to the visible bands and its corresponding scores for carrot and celeriac in Figure 4 and Figure 5 do not show variation for the former and consistency for the latter. Furthermore, the pseudo-RGB images in Figure 3 explain the assessors difficulties in scoring the discoloration attribute. That is, the human eye is not able to observe the subtle changes that a vision system can distinguish using both the visible and NIR regimes. This is an important achievement for the spectral imaging system and can be utilized in developing an on-line quality monitoring set-up.

Based on these findings, a simple spectral vision-based system with a few visible and NIR wavelengths can be used for prediction of those sensory attributes with some consistent variation over the batches. In our work, taste and smell sensory attributes fulfill this. However, the number of batches and samples were limited. The use of more batches and samples that cover possible variability of the population might change this. Further research work is required in this case. In addition, the appropriate sensory attributes may also depend on the characteristics of the food item. For example, in (Løkke, Seefeldt, Skov, & Edelenbos, 2013), color and texture were the appropriate sensory attributes to be predicted using multispectral images of green Wild Rocket vegetables. The suitable wavelengths are also different based on the spectral characteristics of the food item.

Another important issue for such analyses is the modeling method for prediction. In all similar previous works (Barbina, ElMasrya , & Sun, 2012; Kamruzzaman , ElMasry , Sun, & Allen, 2013; Løkke, Seefeldt, Skov, & Edelenbos, 2013), the PLSR regression method was used for prediction and the relevant wavelengths were found in a separate step. However, both tasks can be performed in a single step using the sparse EN regression method as also was used in this work. In this way, the number of contributing wavelengths in prediction will be reduced, which is important regarding simplification of an imaging system in practice. Generally, the accuracy of EN method is very good and comparable to the PLSR.

The two main limitations of this work were the number of batches and ranges of wavelengths. To improve the prediction models in the future in terms of generalization, the number of batches should be increased. In this way, the calibration set will cover the most possible with-in and between batch variations of the population. In addition, because of the important effect of the NIR ranges, the vision system range can be extended to higher wavelengths.

5 Conclusion

The relationship between the multispectral images and the sensory evaluations of two types of vegetables was investigated in this paper. Two types of stir-fried vegetables, carrot and celeriac

were analyzed over a period of 14 days. There were two different batches of each vegetable type. The spectral features, formed from multispectral images, were used to develop regression models for prediction of sensory attributes. The results show that the sensory attributes that had some variation over the storage days and consistency over the two batches resulted in better models in terms of generalization. For carrot, the smell and for celeriac, the off-taste were the attributes that gave the best results. Based on this, the use of more batches and further samples can help to develop better prediction models in terms of generalization. In addition, analysis of wavelengths showed that, both visible as well as NIR bands were among the most contributing wavelengths in the image features that were used by the prediction models. However, the discoloration scores were not appropriate due to the limitation in human visual perception. Therefore, we conclude that a vision-based quality assessment system should utilize multispectral images of some visible and NIR wavelengths together with an appropriate set of calibration sensory attributes (in this case excluding color), to improve the prediction task. In addition, the multispectral images provided a basis for assessing color changes not visible to the human eye.

Acknowledgements

This work was (in part) financed by the Center for Imaging Food Quality project which is funded by the Danish Council for Strategic Research (contract no 09-067039) within the Program Commission on Health, Food and Welfare.
In addition, this study was (in part) financed by the Danish Food Industry Agency.
The authors would like to thank Rene Thrane and Peter Reimer Stubbe for carrying out the manual experiments in the laboratory at DTU National Food Institute. Jeannette Møller and Rie Sørensen, DTU National Food Institute are thanked for assisting with sensory analysis. Student Helene Carlsen is thanked for assistance with manual experiments.
The authors also would like to thank Professor Per B. Brockhoff for giving technical comments on sensory analysis.

References

Adler-Nissen, J. (2007). Continuous wok-frying of vegetables: process parameters influencing scale up and product quality. *Journal of Food Engineering, 83(1)*, 54-60.

Adler-Nissen, J., Akkerman, R., Frosch, S., Grunow, M., Løje, H., Risum, J., . . . Johansson, G. Ø. (2013). Improving the Supply Chain and Food Quality of Professionally Prepared Meals. *Trends in Food Science & Technology, 29(1)*, 74-79.

Adler-Nissen, J., Akkerman, R., Frosch, S., Grunow, M., Løje, H., Risum, J., . . . Johansson, G. Ø. (2013). Improving the Supply Chain and Food Quality of Professionally Prepared Meals. *Trends in Food Science & Technology, 29(1)*, 74-79.

Andrés, S., Murray, I., Navajas, E. A., Fisher, A. V., Lambe, N. R., & Bünger, L. (2007). Prediction of sensory characteristics of lamb meat samples by near infrared reflectance spectroscopy. *Meat Science, 76*, 509–516.

Andresen, M. S., Dissing, B. S., & Løje, H. (2013). Quality assessment of butter cookies applying multi-spectral imaging. *Food Science and Nutrition, 1*, 315-323.

Baiano, A., Terracone, C., Peri, G., & Romaniello, R. (2012). Application of hyperspectral imaging for prediction of physico-chemical and sensory characteristics of table grapes. *Computers and Electronics in Agriculture, 87*, 142-151.

Baiano, A., Terracone, C., Peri, G., & Romaniello, R. (2012). Application of hyperspectral imaging for prediction of physico-chemical and sensory characteristics of table grapes. *Computers and Electronics in Agriculture, 87*, 142-151.

Barbina, D. F., ElMasrya , G., & Sun, D.-W. (2012). Predicting quality and sensory attributes of pork using near-infrared hyperspectral imaging. *Analytica Chimica Acta, 719*, 30– 42.

Barni, M., Cappellini, V., & Mecocci, A. (1995). A vision system for automatic inspection of meat quality. *Image Analysis and Processing, Lecture Notes in Computer Science, 974*, 748-753.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J. Royal Statistical Society, 57*, 289-300.

Breiman, L. (2001). Random Forests. *Machine Learning, 45(1)*, 5–32.

Carstensen, J. M., Hansen, M. E., Lassen, N. K., & Hansen, P. W. (2006). Creating surface chemistry maps using multispectral vision technology. *9th Medical image computing and computer assited invention (MICCAI) – Workshop on biophotonics imaging for diagnostics and treatment*, *17.* Lyngby, Denmark.

Christensen , R. (2012). *ordinal: Regression Models for Ordinal Data*. Retrieved from http://www.cran.r-project.org/package=ordinal/

Christensen, R. H. (2015 - a, January 21). Analysis of ordinal data with cumulative link models estimation with the R-package ordinal. Copenhagen, Denmark.

Christensen, R. H. (2015 - b, January 21). A Tutorial on fitting Cumulative Link Mixed Models with clmm2 from the ordinal Package. Copenhagen, Denmark.

Clemmensen, L. H., Dissing, B. S., Hyldig, G., & Løje, H. (2012). Multispectral imaging of wok-fried vegetables. *Journal of Imaging Science and Technology, 56(2)*, 20404-1e20404-6.

Dabbaghchian, S., Ghaemmaghami, M. P., & Aghagolza, A. (2010). Feature extraction using discrete cosine transform and discrimination power analysis with a face recognition technology. *Pattern Recognition, 43*, 1431-1440.

Daugaard, S. B., Adler Nissen, J., & Carstensen, J. M. (2010). New vision technology for multidimensional quality monitoring of continuous frying of meat. *Food Control, 21*, 626-632.

Dissing, B. S., Clemmensen, L. H., Løje, H., Ersbøll, B. K., & Adler Nissen, J. (2009). Temporal reflectance changes in vegetables. *Proceedings for IEEE Color and Reflectance in Imaging and Computer Vision Workshop.* Kyoto, Japan.

Dissing, B. S., Nielsen, M. E., Ersbøll, B. K., & Frosch, S. (2011). Multispectral imaging for determination of astaxanthin concentration in salmonids. *Plos one 6, 5*, e19032.

Dudoit, S., Shaffer, J. P., & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci., 18*(1), 71-103.

ElMasry, G., & Sun, D.-W. (2010). principles of hyperspectral imaging technology. In D.-W. Sun, *Hyperspectral Imaging for Food Quality Analysis and Control,.* San diego: Academic Press.

ElMasry, G., Iqbal, A., Sun, D. -W., Allen, P., & Ward, P. (2011). Quality classification of cooked, sliced turkey hams using NIR hyperspectral imaging system. *Journal of Food Engineering, 103*(3), 333–344.

Garrido-Novell, C., Pérez-Marin, D., Amigo, J. M., Fernández-Novales, J., Guerrero, J. E., & Garrido-Varo, A. (2012). Grading and color evolution of apples using RGB and hyperspectral imaging. *Journal of Food Engineering, 113*(2), 281-288.

Gonzalez, R. C., & Woods, R. E. (2001). *Digital Image Processing.* New Jersey: Prentice Hall.

Gundersen, H. (2002). The smooth fractionator. *JMicrosc, 207*, 191–210.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning.* Springer.

Herold, B., Kawano, S., Sumpf, B., Tillmann, P., & Walsh, K. B. (2008). VIS/NIR spectroscopy. In M. Zude, *Optical Monitoring of Fresh and Processed Agricultural Crops.* CRC Press.

Hyldig, G., & Green-Petersen, D. M. (2004). Quality Index Method—An Objective Tool for Determination of Sensory Quality. *Journal of Aquatic Food Product Technology, 13*, 71-80.

Hyldig, G., Martinsdottir, E., Sveinsdottir, K., Schelvis, R., & Bremner, A. (2010). Quality Index Methods. In N. L. Fidel Toldra (Ed.), *Sensory Analysis of Foods of Animal Origin.* Taylor & Francis.

ISO-8586-1. (1993). Sensory analysis – a general guidance for the selection, training and monitoring of assessors. Part 1: Selected assessors. *International Standard*.

ISO-8589. (2010). Sensory analysis – General guidance for the design of test rooms. *International Standard*.

Jain, A. K. (1989). *Fundamentals of Digital Image Processing.* Englewood Cliffsy, New Jers: Prentice Hall.

Kamruzzaman , M., ElMasry , G., Sun, D. W., & Allen, P. (2013). Non-destructive assessment of instrumental and sensory tenderness of lamb meat using NIR hyperspectral imaging. *Food Chemistry, 341*, 389-396.

Larsen, A. L., Hviid, M. S., Jørgensen, M. E., Larsen, R., & Dahl, A. L. (2014). Vision-based method for tracking meat cuts in slaughterhouses. *Meat Science, 96*, 366–372.

Lawless, H. T., & Heymann, H. (1999). *Sensory Evaluation of Food: Principles and Practices.* Springer.

Løkke, M. M., Seefeldt, H. F., Skov, T., & Edelenbos, M. (2013). Color and textural quality of packaged wild rocket measured by multispectral. *Postharvest Biology and Technology, 75*, 86–95.

Martínez Vega , M. V., Sharifzadeh, S., Wulfsohn , D., Skov, T., Clemmensen, L. H., & Toldam-Andersen , T. B. (2013). A sampling approach for predicting the eating quality of apples using visible-near infrared spectroscopy. *Journal of the Science of Food and Agriculture, 93*(15), 3710–3719.

Martinsdóttir, E., Schelvis, R., Hyldig, G., & Sveinsdóttir, K. (2009). Sensory evaluation of seafood: general principles and guidelines. In H. Rehbein, & J. Oehlenschlager, *Fishery Products: Quality, Safety and Authenticity* (pp. 411-424). Chichester: Wiley-Blackwell.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *J. IEEE Transactions, 9*, 62-66.

Pallottino, F., Menesatti, P., Costa, C., Paglia, G., De Salvador, F. R., & Lolletti, D. (2010). Image analysis techniques for automated hazelnut peeling. *3*(1), 155-159.

Prieto, N., Andrés, S., Giráldez, F. J., Mantecón, A. R., & Lavin, P. (2006). Potential use of near infrared reflectance spectroscopy (NIRS) for the estimation of chemical composition of oxen meat samples. *Meat Science, 74*(3), 487–496.

Sharifzadeh, S., Clemmensen, L. H., Borggaard, C., Støier, S., & Ersbøll, B. K. (2014). Supervised feature selection for linear and non-linear regression of L* a* b* color from multispectral images of meat. *Engineering Applications of Artificial Intelligence, 27*, 211-227.

Sharifzadeh, S., Clemmensen, L. H., Løje, H., & Ersbøll, B. K. (2013). Statistical Quality Assessment of Pre-fried Carrots Using Multispectral Imaging. *Lecture Notes in Computer Science, 7944*, 620-629.

Skibinski, O. D., Diz, P. A., & Carvajal-Rodríguez, A. (2011, March). Multiple hypothesis testing in proteomics: a strategy for experimental work. *Molecular and Cellular Proteomics, 10*(3).

Sun , D.-W. (2009). *Infrared Spectroscopy for Food Quality Analysis and Control.* Elsevier.

Taghizadeh, M., Gowen, A. A., & O'Donnell, C. P. (2011). The potential of visible-near infrared hyperspectral imaging to discriminate between casing soil, enzymatic browning and undamaged tissue on mushroom (Agaricus bisporus) surfaces. *Computers and Electronics in Agriculture, 77*(1), 74–80.

Tsenkova, R., Atanassova, S., Toyoda, K., Ozaki, Y., Itoh, K., & Fearn, T. (1999). Near-Infrared Spectroscopy for Dairy Management: Measurement of Unhomogenized Milk Composition. *Journal of Dairy Science, 82*(11), 2344–2351.

Tutz, G., & Hennevogl, W. (1996). Random effects in ordinal regression models. *Computational Statistics & Data Analysis, 22*, 537-557.

Varela, P., & Ares, G. (2012). Sensory profiling, the blurred line between sensory and consumer science. A review of novel methods for product characterization. *FOOD RESEARCH INTERNATIONAL, 48*(2), 893-908. doi:10.1016/j.foodres.2012.06.037

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Royal Statistics, 67(2)*, 301-320.

Appendix

A.1 QIM test

One aspect of Sensory testing includes the descriptive and discriminative tests to measure the intrinsic quality of a product such as taste, appearance, etc. On the other hand, it can also include the consumer attitude and emotional response toward the product including both intrinsic and extrinsic quality of the product such as price, origin and etc. The QIM is a sensory evaluation method that was originally developed for fish species. The aim of developing this method was to integrate both of the above mentioned aspects of the sensory analysis. It can build a bridge between research, product development, industry, marketing personnel and consumers (Hyldig & Green-Petersen, 2004). It is a structured scaling method with a scoring system from 0 to a maximum value demerit points. A food item is inspected and the fitting demerit point is recorded. The scores for all the attributes are then summed to give an overall sensory score, the so-called quality index. QIM gives scores of zero for very fresh food while increasingly larger totals result as when the food deteriorates.

A.2 Multiple hypothesis testing

Feature Assessment based on multiple hypothesis testing is a statistical approach used for test and selection of features in problems that the number of features are very high compared to the number of observations $N \ll P$. It is used mostly used for genomic data (Skibinski, Diz, & Carvajal-Rodríguez, 2011; Dudoit, Shaffer, & Boldrick, 2003) to assess the significance of individual features (genes).

Considering having M features and their p-value (e. g. by using the theoretical t-distribution probabilities, which assumes the features are normally distributed or a permutation distribution that does not make any assumption about their distribution), a hypothesis H is formed so that:

$$\begin{cases} H = 0 & Negative(Null) \\ H = 1 & Positive \end{cases} \quad \text{A. 1}$$

This hypothesis is tested for all features $j = 1,2,\dots,M$ and it is accepted $H_j = 1$ or in other words the result is significant at level α if $p_j < \alpha$. This test has $type\ I$ error equal to α (for each individual test). That is, the probability of falsely rejecting $H_j = 0$ is α as shown in table A.1.

Table A.1: Possible outcomes from $\boldsymbol{M}$ hypothesis tests.

	Called Not Significant	**Called Significant**	**Total**
$\boldsymbol{H = 0}$	U	V (Type I error)	M_0
$\boldsymbol{H = 1}$	T (Type II error)	S	M_1
Total	$M - R$	R	M

Since there are a lot of individual tests (M is high), the overall measure of this error is quite high and should be corrected. In cases that M is quit high and high correlation exists between the co-variates,

The Benjamin-Hachberg (BH) (Benjamini & Hochberg, 1995) method is used. In this method the False Discovery Rate (FDR) is introduced as follows:

$$\mathrm{FDR} = \mathrm{E}\left(\frac{\mathrm{V}}{\mathrm{R}}\right) \quad \text{A. 2}$$

It is the expected proportion of the false positive features V among the R features that are called significant. In this method, the FDR rate is bounded by a user defined level α. It is calculated based on the p-values obtained from an asymptotic approximation of the test statistic like a Gaussian or a permutation distribution. If the hypotheses are independent, Benjamini and Hochberg (Benjamini & Hochberg, 1995) showed that regardless of how many null hypotheses are true and regardless of the distribution of the p-values when the null hypothesis is false $H = 1$, this procedure has the property:

$$FDR \leq \alpha \qquad \text{A. 3}$$

In this method, the FDR is fixed at α level and the p-values are ordered $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(M)}$. Then a threshold point (L) is defined based on a threshold line $\alpha\frac{j}{M}, j = 1,2, \ldots, M$ so that:

$$L = \max\left\{j: p_j < \alpha\frac{j}{M}\right\} \qquad \text{A. 4}$$

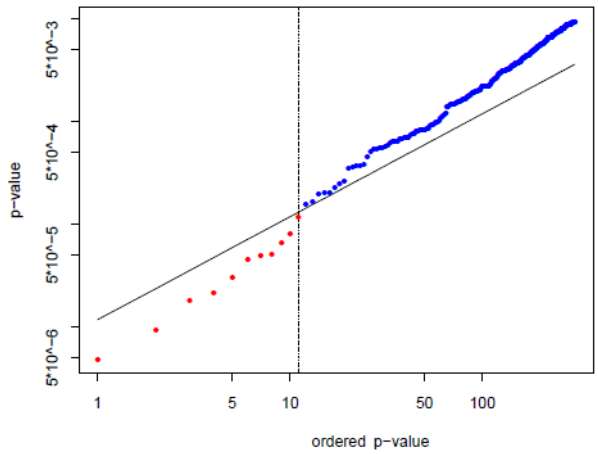


Figure A.1: A plot of the ordered p-values $\boldsymbol{p}_{(j)}$, the threshold line ($\boldsymbol{\alpha\frac{j}{M}}$) as well as the critical point of the BH method (Hastie, Tibshirani, & Friedman, 2009).

This is illustrated in Figure A.1. The null hypotheses is rejected ($H = 1$) for all tests for which $p_j \leq p_{(L)}$, the BH rejection threshold (Hastie, Tibshirani, & Friedman, 2009). The red points in figure A.1 are the significant points that the null hypothesis is rejected for them ($H = 1$). As the FDR rate was kept fixed, the $type\ I$ error is limited.

Bibliography

A. d'Aspremont, F. B. and Ghaoui, L. E. (2007). Optimal solutions for sparse principal component analysis. *CoRR*, abs/0707.0705.

A. d'Aspremont, F. B. and Ghaoui, L. E. (2008). Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9:1269–1294.

Adler-Nissen, J. (2007). Continuous wok-frying of vegetables: Process parameters influencing scale up and product quality. *Journal of Food Engineering*, 83(1):54–60.

Aguilera, J. M. (2005). Why food microstructure? *Journal of Food Engineering*, 67:3–11.

Ahmed, N., Natarajan, T., and Rao, K. R. (1974). Discrete cosine transfom. *IEEE Trans. Comput.*, 23(1):90–93.

Alelyani, S., Tang, J., and Liu, H. (2013). Feature selection for clustering: a review. *Data Clustering: Algorithms and Applications*, pages 29–60.

Andresen, M. S., Dissing, B. S., and Løje, H. (2013). Quality assessment of butter cookies applying multispectral imaging. *Food Science & Nutrition*, 1(4):315–323.

B. Moghaddam, Y. Weiss, S. A. (2006). Spectral bounds for sparse pca: exact and greedy algorithms. In *In proc. NIPS*, pages 915 – 922.

Bair, E., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, 101:119–137.

Balaban, M. O. and Odabasi, A. Z. (2006). Measuring color with machine vision. *Food technology*, 60(12):32 – 36.

Barshan, E., Ghodsi, A., Azimifar, Z., and Zolghadri Jahromi, M. (2011). Supervised principal component analysis: visualization, classification and regression on subspaces and submanifolds. *Pattern Recogn.*, 44(7):1357–1371.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bernd Herold, Sumio Kawano, B. S. P. T. K. B. W. (2008). *VIS/NIR Spectroscopy*. CRC Press.

Bishop, C. M. (2003). Bayesian regression and classification. In *Advances in Learning Theory: Methods, Models and Applications*, pages 267–285. IOS Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning (Information science and statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Bouman, C. A. (1997). Cluster: an unsupervised algorithm for modeling Gaussian mixtures. Available from http://engineering.purdue.edu/~bouman.

Bourne, M. (2002). *Food texture and viscosity: concept and measurement*. Academic Press.

Bouvrie, J. V., Ezzat, T., and Poggio, T. (2008). Localized spectro-temporal cepstral analysis of speech. In *ICASSP*, pages 4733–4736. IEEE.

Brereton, R. (2009). *Chemometrics for Pattern Recognition*. Wiley.

Brewer, M. S., Novakofski, J., and Freise, K. (2006). Instrumental evaluation of ph effects on ability of pork chops to bloom. *Meat Science*, 72(4):pp. 596–602.

Cabassi, G., Profaizer, M., Marinoni, L., Rizzic, N., and Cattaneod, T. M. (2013). Estimation of fat globule size distribution in milk using an inverse light scattering model in the near infrared region. *Journal of Near Infrared Spectroscopy*, 21(5):359 – 373.

Cadima, J. and Jolliffe, I. T. (1995). Loading and correlations in the interpretation of principle compenents. *Journal of Applied Statistics*, 22(2):203–214.

Cao, C. and Jun, Q. (2008). Study on color space conversion between CMYK and CIE L*a*b* based on generalized regression neural network. volume 2, pages pp. 275–277.

Cao, C. and Jun, Q. (2011). Study on color space conversion based on RBF neural network. *Advanced Materials Research*, 174(28):pp. 28–31.

Ceccarelli, M., d'Acierno, A., and Facchiano, A. M. (2009). A scale space approach for unsupervised feature selection in mass spectra classification for ovarian cancer detection. *BMC Bioinformatics*, 10(S-12):9.

Chang, C. C. and Lin, C. J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2.

Chang, H. and Yeung, D.-Y. (2006). Locally linear metric adaptation with application to semi-supervised clustering and image retrieval. *Pattern Recogn.*, 39(7):1253–1264.

Chun, H. and Keles, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society, Series B Stat Methodol*, pages 3–25.

Clemmensen, L. H. (2010). *Data analysis in high-dimensional sparse spaces.* PhD thesis.

Croux, C., Filzmoser, P., and Fritz, H. (2013). Robust sparse principal component analysis. *Technometrics*, 55(2):202–214.

Dabbaghchian, S., Ghaemmaghami, M. P., and Aghagolzadeh, A. (2010). Feature extraction using discrete cosine transform and discrimination power analysis with a face recognition technology. *Pattern Recognition*, 43(4):1431 – 1440.

Dash, M., Choi, K., Scheuermann, P., and Liu, H. (2002). Feature selection for clustering - a filter solution. In *In Proceedings of the Second International Conference on Data Mining*, pages 115–122.

d'Aspremont, A., El Ghaoui, L., Jordan, M., and Lanckriet, G. (2007). A direct formulation for sparse pca using semidefinite programming. *SIAM Review*, 49(3):434–448.

de Jong, S. (1993). Simpls: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251 – 263.

Dissing, B. S. (2011). *New vision technology for multidimensional quality monitoring of food processes.* PhD thesis, Technical University of Denmark, Informatics and Mathematical Modelling.

Dissing, B. S., Carstensen, J. M., and Larsen, R. (2010). Multispectral colormapping using penalized least square. *Journal of Imaging Science and Technology*, 54(3):pp. 0304011–16.

Dissing, B. S., Clemmensen, H. L., Løje, H., Ersbøll, K. B., and Nissen, J. A. (2009). Temporal reflectance changes in vegetables. pages pp. 1917–1922.

Diz A. P., Carvajal-Rodríguez A., S. D. O. (2011). Multiple hypothesis testing in proteomics: a strategy for experimental work. *Molecular and Cellular Proteomics*, 10(3).

Du, C.-J. and Sun, D.-W. (2004). Recent developments in the applications of image processing techniques for food quality evaluation. *Trends in Food Science and Technology*, 15(5):230 – 249.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern classification (2Nd Edition)*. Wiley-Interscience.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.*, 18(1):71–103.

Dy, J. G. and Brodley, C. E. (2000). Feature subset selection and order identification for unsupervised learning. In *In Proc. 17th International Conf. on Machine Learning*, pages 247–254. Morgan Kaufmann.

Dy, J. G. and Brodley, C. E. (2004). Feature selection for unsupervised learning. *J. Mach. Learn. Res.*, 5:845–889.

ElMasry, G. and Sun, D.-W. (2010). {CHAPTER} 1 - principles of hyperspectral imaging technology. In Sun, D.-W., editor, *Hyperspectral Imaging for Food Quality Analysis and Control*, pages 3 – 43. Academic Press, San Diego.

Fayyad, U., Reina, C., and Bradley, P. (1998). Initialization of iterative refinement clustering algorithms. In *Proc. of KDD-1998*, pages 194–198. AAAI Press.

Fdhal, N., Kyan, M., Androutsos, D., and Sharma, A. (2009). Color space transformation from RGB to CIELAB using neural networks. volume Springer, pages pp. 1011–1017.

Fu, J., Lee, S., Wong, S., Yeh, J., Wang, A., and Wu, H. (2005). Image segmentation feature selection and pattern classification for mammographic microcalcifications. *Computerized Medical Imaging and Graphics*, 29(6):419 – 429.

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *J. Mach. Learn. Res.*, 5:73–99.

Gamal, E., Ning, W., and Clément, V. (2009). Detecting chilling injury in red delicious apple using hyperspectral imaging and neural networks. *Postharvest Biology and Technology*, 52(1):pp. 1–8.

Godtliebsen, F., Marron, J. S., and Chaudhuri, P. (2002). Significance in scale space for bivariate density estimation. *Journal of Computational and Graphical Statistics*, 11(1):1–21.

Goluband D. K. Slonim, T. R., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., and M. A. Caligiuri, C.D. Bloomfield, E. L. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, pages 531 – 537.

Gomez, D. D., Clemmensen, L. H., Ersbøll, B. K., and Carstensen, J. M. (2007). Precise acquisition and unsupervised segmentation of multi-spectral images. *Computer Vision and Image Understanding*, 106(2–3):183 – 193. Special issue on Advances in Vision Algorithms and Systems beyond the Visible Spectrum.

Gonzalez, R. C. and Woods, R. E. (2001). *Digital image processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2nd edition.

Goodman, J. W. (2007). *Speckle Phenomena in Optics: Theory and Applications*.

Gretton, A., Bousquet, O., Smola, A., and Sch:olkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings Algorithmic Learning Theory*, pages 63–77. Springer-Verlag.

Gundersen, H. J. (2002). The smooth fractionator. *Journal of microscopy*, pages 191–210.

Hagan, M. T., Demuth, H. B., and Beale, M. (1996). *Neural network design*. PWS Publishing Co., Boston, MA, USA.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, New York. Autres impressions : 2011 (corr.), 2013 (7e corr.).

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67.

Jansen, M. (2001). *Noise reduction by wavelet thresholding*. Lecture notes in statistics. Springer, New York.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547.

Journée, M., Nesterov, Y., Richtárik, P., and Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11:517–553.

Kamruzzaman, M., ElMasry, G., Sun, D.-W., and Allen, P. (2013). Non-destructive assessment of instrumental and sensory tenderness of lamb meat using {NIR} hyperspectral imaging. *Food Chemistry*, 141(1):389 – 396.

Kittler, J. and Illingworth, J. (1986). Minimum error thresholding. *Pattern Recognition*, 19(1):41 – 47.

Kolenda, T., Sigurdsson, S., Winther, O., Hansen, L. K., and Larsen, J. (2002a). *DTU:Toolbox*. ISP Group, Informatics and Mathematical Modeling, Technical University of Denmark.

Kolenda, T., Sigurdsson, S., Winther, O., Hansen, L. K., and Larsen, J. (2002b). Dtu:toolbox.

konicaminolta (2015). CR-400 Minolta. Available online `http://www.konicaminolta.eu/en/measuring-instruments/products/colour-measurement/chroma-meters/cr-400-410/introduction.html`(accessed 9-May-2015).

Krier, C., François, D., Rossi, F., and Verleysen, M. (2007). Feature clustering and mutual information for the selection of variables in spectral data. In *ESANN 2007, 15th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 25-27, 2007, Proceedings*, pages 157–162.

Larrain, R. E., Schaefer, D. M., and Reed, J. D. (2008). Use of digital images to estimate CIE color coordinates of beef. *Food Food Research International*, 41:pp. 380–385.

Larsen, A. B. L., Hviid, M. S., Jørgensen, M. E., Larsen, R., and Dahl, A. L. (2014). Vision-based method for tracking meat cuts in slaughterhouses. *Meat Science*, 96(1):366 – 372.

Lavielle, M. (1999). Detection of multiple changes in a sequence of dependent variables. *Stochastic Processes and their Applications*, 83(1):79 – 102.

León, K., Mery, D., Pedreschi, F., and J., L. (2006). Color measurement in L*a*b* units from RGB digital images. *Elsevier-Food research International*, 39(10):pp. 1084–1091.

Lindeberg, T. (1991). *Discrete scale-space theory and the scale-space primal Sketch*. PhD thesis, Royal Inst. of Technology, Stockholm, Sweden.

Lindeberg, T. (1996). Scale-space: a framework for handling image structures at multiple scales.

Ljungqvist, M., Ersboll, B., Kobayashi, K., Nakauchi, S., Frosch, S., and Nielsen, M. (2012). Near-infrared hyper-spectral image analysis of astaxanthin concentration in fish feed coating. In *Imaging Systems and Techniques (IST), 2012 IEEE International Conference on*, pages 136–141.

Løkke, M. M., Seefeldt, H. F., Skov, T., and Edelenbos, M. (2013). Color and textural quality of 556 packaged wild rocket measured by multispectral imaging. *Postharvest Biology and Technology*, 75:86 – 95.

Lu, Z. and Zhang, Y. (2009). An augmented lagrangian approach for sparse principal component analysis.

Martelli, F., Del Bianco, S., Ismaelli, A., and Zaccanti, G. (2010). *Light propagation through biological tissue and other diffusive media*. SPIE.

Mateo, M. J., O'Callaghan, D. J., Everard, C. D., Castillo, M., Payne, F. A., and O'Donnell, C. P. (2010). Evaluation of on-line optical sensing techniques for monitoring curd moisture content and solids in whey during syneresis. *Food Research International*, 43(1):177 – 182.

Mendoza, F., Dejmekb, P., and Aguilera, J. M. (2006). Calibrated color measurements of agricultural foods using image analysis. *Postharvest Biology-Elsevier Science B.V.*, 4:pp. 285–295.

Meng, D., Zhao, Q., and Xu, Z. (2012). Improve robustness of sparse {PCA} by l1-norm maximization. *Pattern Recognition*, 45(1):487 – 497.

Michelsburg, M., Le, T.-T., Vieth, K.-U., Längle, T., Struck, G., and Puente León, F. (2012). From experiments to realizations: hyper-spectral systems. In *Sensor Based Sorting*, Aachen.

Moghaddam, B., Weiss, Y., and Avidan, S. (2006). *Spectral Bounds for Sparse PCA: Cact and Greedy Algorithms*. MIT Press.

Nielsen, O. H. A., Dahl, A. L., Larsen, R., Møller, F., Nielsen, F. D., Thomsen, C. L., Aanæs, H., and Carstensen, J. M. (2011a). Supercontinuum light sources for hyperspectral subsurface laser scattering - applications for food inspection. In *Image Analysis - 17th Scandinavian Conference, SCIA 2011, Ystad, Sweden, May 2011. Proceedings*, pages 327–337.

Nielsen, O. H. A., Dahl, A. L., Rasmus, L., Flemming, M., Nielsen, F. D., Thomsen, C. L., Henrik, A., and Carstensen, J. M. (2011b). *In depth analysis of food structures: hyperspectral subsurface laser scattering*, pages 29–34. Technical University of Denmark.

Pallottino, F., Menesatti, P., Costa, C., Paglia, G., De Salvador, F., and Lolletti, D. (2010). Image analysis techniques for automated hazelnut peeling determination. *Food and Bioprocess Technology*, 3(1):155–159.

Papandreou, G. and Maragos, P. (2005). Image denoising in nonlinear scale-spaces: automatic scale selection via cross-validation. In *Proceedings of the 2005 International Conference on Image Processing, ICIP 2005, Genoa, Italy, September 11-14, 2005*, pages 481–484.

Pearson, K. (1901). On lines and planes of closest fit to systems of points is space. *Philosophical Magazine Series 6*, 2(11):559–572.

Prigent, S., Descombes, X., Zugaj, D., and Zerubia, J. (2010). Spectral analysis and unsupervised svm classification for skin hyper-pigmentation classification. In *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2010 2nd Workshop on*, pages 1–4.

Ramirez, M. A. and Minami, M. (2003). *Low-bit-rate speech coding*. John Wiley & Sons, Inc.

Roth, V. and Lange., T. (2003). Feature selection in clustering problems. In *Advances in neural information processing systems*.

Sánchez, N. H., Lurol, S., Roger, J., and Bellon-Maurel, V. (2003). Robustness of models based on nir spectra for sugar content prediction in apples. *Journal of Near Infrared Spectroscopy*, 11(2):97–107.

Schlenke, J., Hildebrand, L., Moros, J., and Laserna, J. J. (2012). Adaptive approach for variable noise suppression on laser-induced breakdown spectroscopy responses using stationary wavelet transform. *Analytica Chimica Acta*, 754(0):8 – 19.

Serpico, S. and Moser, G. (2007). Extraction of spectral channels from hyperspectral images for classification purposes. *Geoscience and Remote Sensing, IEEE Transactions on*, 45(2):484–495.

Sharifzadeh, S., Clemmensen, L., Ersbøll, B., and Vega, M. (2013a). Optimal vision system design for characterization of apples using us/vis/nir spectroscopy data. In *Systems, Signals and Image Processing (IWSSIP), 2013 20th International Conference on*, pages 11–14.

Sharifzadeh, S., Clemmensen, L. H., Borggaard, C., Støier, S., and Ersbøll, B. K. (2014). Supervised feature selection for linear and non-linear regression of l*a*b* color from multispectral images of meat. *Engineering Applications of Artificial Intelligence*, 27(0):211 – 227.

Sharifzadeh, S., Clemmensen, L. H., Løje, H., and Ersbøll, K. B. (2013b). Statistical quality assessment of pre-fried carrots using multispectral imaging. volume Springer Lecture Notes in Computer Science, pages pp. 620–629.

Sharifzadeh, S., Serrano, J., and Carrabina, J. (2012a). Spectro-temporal analysis of speech for spanish phoneme recognition. In *Systems, Signals and Image Processing (IWSSIP), 2012 19th International Conference on*, pages 548–551.

Sharifzadeh, S., Skytte, J. L., Nielsen, O. H. A., Ersbøll, K. B., and Clemmensen, L. H. (2012b). Regression and sparse regression methods for viscosity estimation of acid milk from it's sls features. pages pp. 52 – 55.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivar. Anal.*, 99(6):1015–1034.

Shi, L. and Xu, L. (2006). Comparative investigation on dimension reduction and regression in three layer feed-forward neural network. volume 4131/2006, pages 51–60.

Sigg, C. D. and Buhmann, J. M. (2008). Expectation-maximization for sparse and non-negative pca. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 960–967. ACM.

Skytte, J. L., Larsen, R., and Dahl, A. B. (2014). *2D Static light scattering for dairy based applications*. PhD thesis.

Specht, D. F. (1993). The general regression neural network-rediscovered. *Neural Netw.*, 6(7):1033–1034.

Specht, F. D. (1991). A general regression neural network. *IEEE Transaction on Neural Networks*, 2(6):568–576.

Sun, D.-W. (2009). *Infrared spectroscopy for food quality analysis and control*. Academic Press.

Sun, D. W. (2010). *Hyperspectral imaging for food quality analysis and control*. Elsevier, London-UK.

Tarn, D., Arianna, C., Inge, K., and P., W. M. (2008). Feature significance for multivariate kernel density estimation. *Computational Statistics & Data Analysis*, 52(9):4225 – 4242.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., and Le, Q. T. (2004). Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics*, 20(17):3034–3044.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, pages 91–108.

Tkalčič, M. and Tasič, J. F. (2003). Color spaces - perceptual, historical and applicational background. *EUROCON- Computer as a Tool*, pages pp. 304 – 308.

Torkkola, K. (2003). Feature extraction by non parametric mutual information maximization. *J. Mach. Learn. Res.*, 3:1415–1438.

Upton, S. (2006). Delta E: the color difference. http://www.colorwiki.com/wiki/Delta_E:_The_Color_Difference. Online; accessed 17-July-2012.

Varela, P. and Ares, G. (2012). Sensory profiling, the blurred line between sensory and consumer science. a review of novel methods for product characterization. *Food Research International*, 48(2):893 – 908.

Vu, V. Q., Cho, J., Lei, J., and Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 26*, pages 2670–2678.

Witten, D. M., Hastie, T., and Tibshirani, R. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*.

Wu, D. and Sun, D.-W. (2013). Colour measurements by computer vision for food quality control: A review. *Trends in Food Science and Technology*, 29(1):5 – 20.

X-Rite (2004). The color guide and glossary. http://graphics.tech.uh.edu/courses/4373/materials/X-Rite_Color_Guide_2004.pdf. Online; accessed 17-July-2012.

Yeung, D.-Y. and Chang, H. (2006). Extending the relevant component analysis algorithm for metric learning using both positive and negative equivalence constraints. *Pattern Recognition*, 39(5):1007 – 1010.

Zelnik-manor, L. and Perona, P. (2004). Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*, pages 1601–1608. MIT Press.

Zhang, Y. and Ghaoui, L. E. (2011). Large-scale sparse principal component analysis with application to text data. In *Advances in Neural Information Processing Systems (NIPS)*.

Zhao, Z., Wang, L., and Liu, H. (2010). Efficient spectral feature selection with minimum redundancy. In Fox, M. and Poole, D., editors, *AAAI*. AAAI Press.

Zhaoran Wang, Huanran Lu, H. L. (2014). Tighten after relax: Minimax-optimal sparse pca in polynomial time. In *NIPS 2014*.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2004). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15:2006.